

(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES  
PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

(19) Weltorganisation für geistiges Eigentum  
Internationales Büro



(43) Internationales Veröffentlichungsdatum  
25. Mai 2001 (25.05.2001)

PCT

(10) Internationale Veröffentlichungsnummer  
**WO 01/36614 A2**

- (51) Internationale Patentklassifikation<sup>7</sup>: C12N 15/00 9. 82319 Stamberg (DE). GRAF, Marcus [DE/DE]; Spiegelgasse 3 b, 93047 Regensburg (DE).
- (21) Internationales Aktenzeichen: PCT/DE00/04073 (74) Anwälte: BETTENHAUSEN, Berthold usw.; Dehmel & Bettenhausen, Müllerstrasse 1, 80469 München (DE).
- (22) Internationales Anmeldedatum: 16. November 2000 (16.11.2000) (81) Bestimmungsstaaten (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.
- (25) Einreichungssprache: Deutsch (84) Bestimmungsstaaten (regional): ARIPO-Patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), eurasisches Patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), europäisches Patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI-Patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).
- (26) Veröffentlichungssprache: Deutsch
- (30) Angaben zur Priorität: 199 55 089.1 16. November 1999 (16.11.1999) DE
- (71) Anmelder (für alle Bestimmungsstaaten mit Ausnahme von US): GENEART GMBH GESELLSCHAFT FÜR ANGEWANDTE BIOTECHNOLOGIE [DE/DE]; Ahornstrasse 12, 93080 Pentling (DE).
- (71) Anmelder und Veröffentlicht:  
(72) Erfinder: SHAO, Yiming [CN/CN]; 27 Nan Wie Avenue, Beijing Xuan Wu Qu 1000050 (CN). — Ohne internationalen Recherchenbericht und erneut zu veröffentlichen nach Erhalt des Berichts.
- (72) Erfinder; und Zur Erklärung der Zweibuchstaben-Codes, und der anderen Abkürzungen wird auf die Erklärungen ("Guidance Notes on Codes and Abbreviations") am Anfang jeder regulären Ausgabe der PCT-Gazette verwiesen.
- (75) Erfinder/Anmelder (nur für US): WAGNER, Ralf [DE/DE]; Franz-von-Taxis-Ring 59, 93949 Regensburg (DE). WOLF, Hans [DE/DE]; Josef Jägerhuberstrasse



(54) Title: THE GENOME OF THE HIV-1 INTER-SUBTYPE (C/B') AND USE THEREOF

(54) Bezeichnung: DAS GENOM DES HIV-1 INTERSUBTYP (C/B') UND SEINE ANWENDUNGEN

(57) Abstract: The invention relates to a polynucleotide, comprising a nucleic acid sequence according to SEQ ID NO:1, 2 or 3, fragments or derivatives thereof, or a polynucleotide hybridised with the nucleic acid sequence according to SEQ ID NO:1, 2 or 3. The invention further relates to polypeptides coded from said nucleotide sequence, or fragment, or derivative of the nucleic acid sequence according to SEQ ID NO:1, 2 or 3. The polynucleotides and polypeptides may be used as medicaments, vaccines or diagnostics, in particular for the treatment, prophylaxis and diagnosis of HIV infections.

**WO 01/36614 A2**

(57) Zusammenfassung: Die vorliegende Erfindung betrifft ein Polynukleotid, umfassend eine Nukleinsäuresequenz gemäß SEQ ID NO:1, 2 oder 3 oder dessen Fragment oder Derivat, oder ein Polynukleotid, das mit der Nukleinsäuresequenz gemäß SEQ ID NO:1, 2 oder 3 hybridisiert. Die vorliegende Erfindung betrifft ferner Polypeptide, kodiert von der Nukleotidsequenz oder Fragment oder Derivat der Nukleotidsequenz gemäß SEQ ID NO: 1, 2 oder 3. Die Polynukleotide und Polypeptide können als Arzneimittel, Impfstoffe oder Diagnostika, insbesondere für die Behandlung, Prävention und Diagnose von HIV-Infektionen, verwendet werden.

## Das Genom des HIV-1 Intersubtyps (C/B) und seine Anwendungen

Die vorliegende Erfindung betrifft ein Polynukleotid, umfassend eine Nukleinsäuresequenz gemäß SEQ ID NO:1, 2 oder 3 oder dessen Fragment oder Derivat, oder ein Polynukleotid, das mit der Nukleinsäuresequenz gemäß SEQ ID NO:1, 2 oder 3 hybridisiert. Die vorliegende Erfindung betrifft ferner Polypeptide, kodiert von der Nukleotidsequenz oder Fragment oder Derivat der Nukleotidsequenz gemäß SEQ ID NO:1, 2 oder 3. Die Polynukleotide und Polypeptide können als Arzneimittel, Impfstoffe oder Diagnostika, insbesondere für die Behandlung, Prävention und Diagnose von HIV-Infektionen, verwendet werden.

In Anbetracht des Ausmaßes und der globalen Verbreitung der durch das humane Immunodefizienz Virus (HIV) verursachten Pandemie mit einer, bis zum Ende dieses Jahrhunderts, geschätzten Anzahl von weltweit mehr als 40 Millionen Infizierten (davon mehr als 90% in Entwicklungsländern) stellt die Entwicklung einer wirksamen HIV-Vakzine eine der größten Herausforderungen an die moderne industrialisierte Welt dar. Bislang wird die Entwicklung eines erfolgreichen HIV-Impfstoffs jedoch noch immer durch die komplizierte Biologie des Virus sowie seine komplexe Interaktion mit dem Immunsystem des Wirtes limitiert. Die wenigen Impfstoff-Kandidaten, die bis zum heutigen Zeitpunkt in Entwicklungsländern in Phase 3-Studien ausgetestet wurden, basierten hauptsächlich auf den externen Glycoproteinen gp120 oder gp160 von HIV Typ-1. Der Ausgang der Studien war jedoch eher enttäuschend: Die Impfstoffe waren nicht nur nicht in der Lage, breit kreuz-neutralisierende Antikörper- und T-Zell-Reaktionen hervorzurufen. Sie konnten nicht einmal Infektionsdurchbrüche verhindern, die bei einigen Impflingen beobachtet worden sind. Einer der Gründe für dieses Versagen liegt sicherlich in den extensiven Sequenzvariationen zwischen den verwendeten Antigenen, welche von laboradaptierten Virusstämmen abstammten, und den genetisch divergenten Viren, welche in den Testregionen (z.B. Thailand) kursierten.

Phylogenetische Analysen der weltweit zirkulierenden HIV-Stämme haben eine Hauptgruppe (M) mit 10 verschiedenen Sequenz-Subtypen (A-J) (Kostrikis *et al.* 1995; Leitner und Albert,

- 1995; Gaywee *et al.* 1996; World Health Organisation Network for HIV Isolation and Characterization, 1994), die im Hüllprotein Sequenzvariationen von bis zu 24 % aufweisen, und außerdem die Viren der O-Gruppe identifiziert, die sich in einigen Leserahmen um mehr als 40 % von den Viren der M-Gruppe unterscheiden (Loussert Ajaka *et al.* 1995; Myers *et al.* 1996;
- 5 Sharp *et al.* 1995; Sharp *et al.* 1999). Zudem entwickelt sich HIV durch die rasche Anhäufung von Mutationen und Intersubtyp-Rekombinationen immer weiter. Unterschiedliche Subtypen, welche innerhalb der Population einer geographischen Region kozirkulieren, stellen die molekulare Grundlage für die Erzeugung und Ausbreitung von gruppenübergreifenden Intersubtyp-Mosaikviren dar. Obwohl die weltweit verbreiteten HIV-1-Varianten durch Serologie
- 10 und Heteroduplex-DNA-Analysen intensiv untersucht wurden, beruhen die meisten phylogenetischen Analysen auf Sequenzen des Hüllproteins, da für viele der prävalenten Subtypen und eine Vielzahl von rekombinanten Formen keine vollsequenzierten Genome vorliegen.
- 15 Für die überwiegende Mehrheit der weltweit neuen HIV-1-Infektionen sind Viren des Subtyps Non-B (also *Nicht-B*-Varianten) verantwortlich. Den Viren des Subtyps C fällt dabei im Hinblick auf die Gesamtzahl von Infizierten sowie der weiten Verbreitung von Neu-Infektionen, insbesondere in Süd-Amerika und Asien, eine herausragende Rolle zu. Auf Grund dessen hat die Charakterisierung von Viren des Subtyps C eine herausragende Priorität für diagnostische,
- 20 therapeutische oder präventive Zwecke.

Mit Ausnahme von Thailand lagen bis vor kurzem nur begrenzte Informationen über die Verteilung und molekulare Charakteristik von in Asien vorkommenden HIV-1-Stämmen vor. Nach Schätzungen der WHO breite sich HIV am schnellsten in Süd- und Südost-Asien aus,

25 welche schon bald die weltweit größte Region mit HIV-Epidemie sein wird. China unterliegt ähnlichen sozialen und ökonomischen Strukturen und unterhält zu diesen Regionen unmittelbare ethnische und wirtschaftliche Verbindungen. In vielen Provinzen Chinas konnte seit Anfang 1995 ein rasanter Anstieg von HIV-Infektionen beobachtet werden. Verglichen mit allen von 1985 bis 1994 dokumentierten 1774 Fällen an HIV und AIDS, wurden im Jahr 1995 alleine

30 schon 1421 Fälle und im Jahr 1997 mehr als 4000 Fälle nachgewiesen. Die WHO geht von mehr als 400.000 HIV-Infektionen in China bis Ende 1997 aus, mit bis dahin 6400 Todesfällen und einer geschätzten Anzahl von 4000 Todesfällen allein im Jahre 1997. Im kürzlich

veröffentlichten nationalen HIV-Molekularepidemiologischen Bericht wurde gefunden, daß die Thai-Stämme des Prototyp-Subtyps B und des Subtyps B' in Yunnan, einer Provinz im Südwesten von China, die angrenzt an das Drogendreieck von Myanmar, Laos und Thailand (Graf et al. 1998), durch Benutzer von Drogen und durch Sammelstellen für kontaminiertes Blut und Plasma bis nach Zentral- und Ost-China verbreitet wurden. In den frühen 90er Jahren wurde dann in die gleiche Region eine zweite Epidemie eingeschleppt, sehr wahrscheinlich durch mit Stämmen des Subtyps C infizierte indische IDUs (intravenous drug user), also Menschen aus Indien, die Drogen intravenös verwenden (Luo et al. 1995; Shao et al. 1999). Innerhalb weniger Jahre verbreiteten sich die Viren des Subtyps C durch Drogenschmuggel schnell in Süd-, Zentral- und sogar in Nordwest-China und verursachten eine weitere Verbreitung der Epidemie innerhalb Chinas. Einem kürzlich veröffentlichten nationalen HIV-Molekularepidemiologischen Untersuchungsbericht zufolge sind fast alle mit Viren des Subtyps C infizierten Personen IDUs und machen damit etwa 40% aller HIV-infizierten IDUs in China aus. Das legt nahe, daß die Viren des Subtyps C zu den wichtigsten Subtypen von HIV-1 zählen, die unter IDUs in China prävalent sind (Shao et al. 1998, Shao et al. 1994).

Dies legt nahe, daß sich die HIV-Epidemie unter den IDUs in China innerhalb weniger Jahre von einem einzelnen vorherrschenden Subtyp (B) auf mindestens 2 vorherrschende Subtypen, B-Thai und C, ausgeweitet hat, was die Möglichkeit der Intersubtyp-Rekombination erhöht. Nach 20 unserem bisherigen Kenntnisstand über Variabilität und Antigenizität unterschiedlicher Virus-Stämme sollten Diagnostika, Therapeutika und Impfstoffe auf regionale Virus-Stämme angepasst sein. Die Anzahl molekularer Reagenzien für Viren des *Nicht-B*-Subtyps sind jedoch noch extrem limitiert. Außer für Viren des Subtyps B oder C sind bislang nur wenige nicht-rekombinante molekulare Klone und wenige Mosaikgenome verfügbar. Was HI-1-Viren des 25 Subtyps C betrifft, sind bislang nur nicht-rekombinante Vertreter und vier A/C-Rekombinannten publiziert, die alle aus Afrika, Süd-Amerika oder Indien stammen (Luo et al. 1995; Gao et al. 1998; Lole et al. 1999). Darüberhinaus beschränken sich die bislang gesammelten Daten über Viren des Subtyps C in China auf genetische Subtypisierungen des env-Gens (Luo et al. 1995; Yu et al. 1997; Salminen et al. 1995).

30 Mehrere klinische Studien zur Bekämpfung von HIV-Infektionen wurden bislang mit Vakzinen durchgeführt. Die enttäuschenden Ergebnisse, die bei klinischen Studien beobachtet wurden,

beinhalteten wiederholt berichtete Infektionsdurchbrüche bei den Impflingen. Dies wurde vor allem auf die umfangreichen Sequenzvariationen zwischen den verabreichten Hüllproteinen und dem infektiösen Input-Virus zurückgeführt, was tatsächlich vorwiegend auf eine unzureichende Charakterisierung der in einer bestimmten geographischen Region zirkulierenden Viruspopulation zurückzuführen ist. Dies resultierte in der Erzeugung von humoralen und - in geringerem Ausmaß - von zellvermittelten Immunantworten gegen virale Antigene, welche nicht relevant waren für die in der Population des Testgebietes zirkulierenden Viren. Zudem konnte gezeigt werden, daß geringe, spezifisch gegen das Hüllprotein gerichtete Antikörper nicht nur keine neutralisierenden Eigenschaften besitzen, sondern darüberhinaus sogar zu einer Verstärkung der Infektion mittels Komplement- oder Fc-Rezeptor beitragen. Des Weiteren erwiesen sich die ausgewählten Antigene und Verabreichungssysteme als extrem schwach für die Induktion der zellvermittelten Immunantwort.

Angesichts eines Mangels an genauer Kenntnis über Subtyp-übergreifende protektive Immunantworten sowie über die komplexe Situation in Entwicklungsländern, wo bekanntermaßen viele Subtypen von HIV-1 kozirkulieren, sollten Impfstoff-Präparationen Mischungen von repräsentativen Antigenen enthalten. Somit besteht also ein Bedarf an der Isolierung und Charakterisierung von Viren des Subtyps C, insbesondere für die Klonierung der kodierenden Region.

Die Aufgabe der vorliegenden Erfindung wird durch den in den Patentansprüchen definierten Gegenstand gelöst.

Die nachfolgenden Figuren erläutern die vorliegende Erfindung.

Figur 1 zeigt eine Darstellung der phylogenetischen Verwandtschaft der das env-Gen C2V3 kodierenden Region des Klons 97cn54 zu den Vertretern der wichtigen Subtypen von HIV-1 (M-Gruppe). "cn-con-c" steht für die env-Konsensussequenz der HIV-1-Stämme des Sub-typs C, welche in China prävalent sind. Der phylogenetische Stammbaum wurde mittels der "neighbour joining"-Methode erstellt. Die Werte an den Knoten geben die "bootstraps" in % an, welche die Eingruppierung rechts unterstützt. Nur "bootstrap"-Werte, die 70% erreichen oder überschreiten, sind angegeben. Die Klammern rechts stellen die Sequenzen der wichtigsten Subtypen von

HIV-1, M-Gruppe, dar.

Figur 2 zeigt eine Darstellung der RIP-Analyse (Recombinant Identification Program), Version 1.3, des gesamten für gagpol kodierenden Bereichs von 97cn54 (Fenstergröße: 200, 5 Schwellenwert für die statistische Signifikanz: 90%, Umgang mit Lücken: STRIP). Die Positionen der offenen Leserahmen von gag und pol sind durch Pfeile im Diagramm oben dargestellt. Die RIP-Analyse basierte auf Hintergrundvergleichen unter Verwendung von Referenzsequenzen, die von ausgewählten Virus-Stämmen abstammten, die die wichtigsten Subtypen von HIV-1 darstellen. Standardvertreter sind durch verschiedene Farben markiert, wie 10 angezeigt. Die x-Achse gibt die Nukleotid-Positionen entlang des Sequenzvergleichs an. Die y-Achse gibt die Ähnlichkeit von 97cn54 mit den aufgelisteten Referenz-Subtypen an.

Figur 3 zeigt eine Darstellung der phylogenetischen Verwandtschaft verschiedener Regionen innerhalb der von 97cn54 abgeleiteten Leserahmen von gagpol mit Standard-Vertretern der 15 wichtigsten Subtypen von HIV-1 (M-Gruppe). Unter Verwendung der "neighbour joining"-Methode basierend auf den folgenden Sequenzabschnitten: (A) Nukleotide 1-478, (B) 479-620, (C) 621-1290, (D) 1291-1830, (E) 1831-2220, (F) 2221-2520 und (G) 2521-2971 wurden phylogenetische Stammbäume erstellt. Die angegebenen Positionen beziehen sich auf das erste Nukleotid des offenen Leserahmens von gag. Graue Bereiche kennzeichnen Cluster der 20 analysierten Sequenzen entweder mit von Subtyp C (A, C, E, G) oder von Subtyp B (B, D, F) abgeleiteten Referenzstämmen. Die Werte an den Knoten geben die "bootstrap"-Werte in Prozent an, durch die das Cluster rechts bestätigt wurde. Es werden nur "bootstrap"-Werte von 70% oder mehr gezeigt.

25 Figur 4 zeigt eine Darstellung der RIP-Analyse, Version 1.3, von verschiedenen Regionen von 97cn54 (Fenstergröße: 200, Schwellenwert für die statistische Signifikanz: 90%, Umgang mit Lücken: STRIP). Die Analyse umfaßte (A) einen Sequenzbereich von 1500 bp Länge vom Startkodon des vif-Gens bis zum 5'-Ende von env einschließlich vif, vpr, dem ersten Exon von tat und rev, vpu und den ersten 200 bp des env-Gens und (B) ein etwa 700 bp langes Fragment, 30 das 300 bp vom 3'-Ende von env, die das komplette nef-Gen und Teile der 3'-LTR-Region umfassen, überlappt. Die Positionen der Startkodons vpr, tat, vpu, env, nef und das 5'-Ende der 3'-LTR-Region sind jeweils oben in den Diagrammen durch Pfeile gekennzeichnet. Die RIP-

- Analyse basierte auf Hintergrund-Vergleichen unter Verwendung von Sequenzen, die von ausgewählten Virusstämmen abgeleitet waren, die die wichtigsten Subtypen von HIV-1 repräsentierten. Die angegebenen Standardvertreter sind durch verschiedene Farben gekennzeichnet. Die x-Achse gibt die Nukleotidpositionen entlang des Sequenzvergleichs an.
- 5 Die y-Achse gibt die Ähnlichkeit von 97cn54 mit den aufgelisteten Referenz-Subtypen an. (C) und (D) zeigen RIP-Analysen von Sequenzen von zwei unabhängigen C-Isolaten (xj24 und xj158) aus China, die das vpr- und vpu-Gen einschließlich des ersten Exons von tat überlappen.

Figur 5 zeigt die Analyse eines phylogenetischen Stammbaums. Phylogenetische Stammbäume  
10 wurden unter Verwendung der "neighbour joining"-Methode erstellt basierend auf (A) einem 380 bp langen Fragment, das 150 bp vom 3'-Ende des vpr-Gens bis zum Ende des vpu-Leserahmens überlappt, (B) den ersten 290 bp der kodierenden Region von nef und (C) auf den 320 bp am 3'-Ende des nef-Gens. Die Werte an den Knoten geben die "bootstrap"-Werte in Prozent an, durch  
15 die das Cluster rechts bestätigt wurde. Es werden nur "bootstrap"-Werte von 70% oder mehr gezeigt. Die Klammern rechts stellen die wichtigsten Subtyp-Sequenzen von HIV-1, Gruppe M, dar.

Figur 6 ist eine schematische Darstellung der mosaikartigen Organisation des Genoms von  
97cn54.

20 Figur 7 ist eine Darstellung des Vergleichs zwischen bekannten und experimentell nachgewiesenen CTL-Epitopen des Prototyps B (HIV-1<sub>LAI</sub>) und den entsprechenden Aminosäure-Sequenzen der Polypeptide gag, pol und env des Stammes 97cn54 vom Subtyp C. Die funktionellen Domänen in GAG (p17 Matrix, p24 Kapsid, p15 Nukleokapsid und Linker-Protein), POL (PR Protease, RT Reverse Transkriptase, IN Integrase) und ENV (gp120 äußeres Glykoprotein, gp41 Transmembranprotein) sind entsprechend bezeichnet. Die Zahlen unterhalb der offenen Leserahmen geben die Aminosäure-Position relativ zu den aminoterminalen Enden der Polypeptide an. Haplotyp-Restriktionen der bekannten CTL-Epitope von HIV-1<sub>LAI</sub> sind am linken bzw. rechten Rand angegeben. Die grünen Balken kennzeichnen Sequenz-Identität  
25 zwischen dem bekannten Epitop und der entsprechenden Sequenz vom Subtyp C, blaue Balken bedeuten 2 oder weniger konservative Fehlpaarungen. Rote Balken stellen vom Subtyp C abgeleitete Sequenz-Bereiche mit mehr als 2 konservativen Fehlpaarungen oder nicht-

konservative Substitutionen im Vergleich zu dem entsprechenden von LAI abgeleiteten Epitop dar.

- Figur 8 zeigt die vollständige kodierende Nukleotidsequenz von 97cn54 von HIV-1, Subtyp C  
5 (SEQ ID NO:1), mit den entsprechenden Aminosäuren im Einbuchstaben-Kode. Alle 3 Leserahmen sind angegeben. Die Sternchen stellen Stopp-Kodons dar.

- Figur 9 zeigt in einer Darstellung das Ergebnis der Aktivitäten zytotoxischer T-Zellen in Milzzellen von BALB/c Mäusen nach intramuskulärer Immunisierung mit den angezeigten  
10 DNA-Plasmiden. Lymphoide Zellen, gewonnen 3 Wochen nach der Primärimmunisierung aus jeweils 5 Mäusen pro Gruppe, wurden mit AMQMLKETI (Einbuchstabencode) Gag-Peptid beladenen syngenen P815 Mastozytom Zellen (bestrahlt mit 20,000 rad) kokultiviert. Kontrollen schlossen Milzzellen nicht-immunisierter Mäuse, stimuliert mit Peptid-beladenen P815 Zellen ein. Zytotoxische Effektor-Zellpopulationen wurden nach einer 5-tägigen Kultur *in vitro* geerntet.  
15 Die zytotoxischen Antworten wurden gegen A20 Zellen, beladen mit dem oben aufgeführte nonameren Peptid oder gegen unbeladene A20 Zellen, in einem Standard <sup>51</sup>Cr Freisetzungstest ausgelesen. Die gezeigten Daten repräsentieren die Mittelwerte aus jeweils Dreifachansätzen. Die ermittelten Standardabweichungen lagen jeweils unter 15% gemessen am Mittelwert.  
20 Die Begriffe "Epitop" oder "Antigene Determinante", wie nachfolgend verwendet, bedeuten eine immunologisch determinante Gruppe eines Antigens, das spezifisch von einem Antikörper erkannt wird. Ein Epitop kann Aminosäuren in räumlicher oder diskontinuierlicher Konformation umfassen und umfaßt mindestens 3, vorzugsweise mindestens 5, Aminosäuren. Ein Epitop kann auch ein einzelnes Segment einer Polypeptid-Kette umfassend eine  
25 kontinuierliche Aminosäure-Sequenz umfassen.

- Der Begriff "Polynukleotid", wie nachfolgend verwendet, bezieht sich auf ein einzel- oder doppelsträngiges Heteropolymer aus Nukleotid-Einheiten beliebiger Länge, wobei diese entweder Ribo- oder Desoxyribonukleotide sein können. Der Begriff umfaßt auch modifizierte  
30 Nukleotide.

- Der Begriff "Derivat", wie nachfolgend verwendet, bezeichnet eine Nukleinsäure, die ebenfalls das oder die Polypeptide kodiert, die von einer anderen Nukleotidsequenz kodiert werden, obwohl sich ihre Nukleotidsequenz von der anderen Nukleotidsequenz unterscheidet. In diesem Sinne bezeichnet der Ausdruck „Derivat“ auch Äquivalente der anderen Nukleotidsequenz, die
- 5 aufgrund der Degeneration des genetischen Codes vorliegen. Unter den Begriff Derivat fallen z.B. Nukleinsäuren, die die gleichen Polypeptide wie die Nukleotidsequenz gemäß SEQ ID NO:1, 2 oder 3 kodieren, aber eine andere Nukleotidsequenz aufweisen, oder es fallen ferner Nukleinsäure-Fragmente unter den Begriff, die das gleiche Polypeptid kodieren wie Nukleinsäure-Fragmente der Nukleotidsequenz gemäß SEQ ID NO:1, 2 oder 3.
- 10 Der Begriff "Polypeptid", wie nachfolgend verwendet, bezieht sich auf eine Kette von mindestens 2 Aminosäure-Resten, die durch Peptidbindungen miteinander verbunden sind. Der Begriff umfaßt daher alle Aminosäure-Ketten, z.B. Oligopeptide und Proteine. Der Begriff bezieht sich auch auf solche Aminosäure-Ketten, bei denen eine oder mehrere Aminosäure(n) 15 modifiziert ist(sind), z.B. durch Acetylierung, Glykosylierung oder Phosphorylierung.
- Der Begriff "kontinuierliche Sequenz" und "Fragmente", wie nachfolgend verwendet, bezieht sich auf einen linearen Abschnitt von Nukleotiden oder Aminosäuren, der von einer Referenz-Sequenz stammt, z.B. von den Sequenzen der vorliegenden Erfindung, wie sie in dem
- 20 Sequenzprotokoll wiedergegeben sind.
- Der Begriff "selektive Hybridisierung" bzw. "selektiv hybridisierbar", wie nachfolgend verwendet, bezieht sich auf Hybridisierungsbedingungen, bei denen zwei Polynukleotide unter stringenten Hybridisierungsbedingungen Duplex-Nukleotidmoleküle bilden. Diese Bedingungen sind im Stand der Technik bekannt und z.B. in Sambrook et al., Molecular Cloning, Cold Spring Harbour Laboratory (1989), ISBN 0-87969-309-6 beschrieben. Beispiele für stringente Hybridisierungsbedingungen sind: (1) Hybridisierung in 4 x SSC bei 65°C oder (2) Hybridisierung in 50% Formamid in 4 x SSC bei 42°C, jeweils gefolgt von mehreren Waschschritten in 0,1 x SSC bei 65°C (1 Stunde lang).
- 30 Der Begriff "viraler Vektor" oder "bakterieller Vektor", wie nachfolgend verwendet, bezieht sich auf gentechnisch veränderte Viren oder Bakterien, mit denen sich die in den SEQ ID NO:1, 2

oder 3 ausgeführten DNA-Sequenzen, davon abgeleitete Derivate, Fragmente, Sequenzen kodierend für Epitope oder Epitop-Strings in unterschiedliche Zellen, bevorzugt in antigenpräsentierende Zellen wie beispielsweise Dendritische Zellen einbringen lassen. Ein bakterieller Vektor kann darüber hinaus geeignet sein, ein von SEQ ID NO:1, 2 oder 3 kodiertes

5 Polypeptid, davon abgeleitete Epitope oder Epitop-Strings direkt zu exprimieren.

Ein Aspekt der vorliegenden Erfindung betrifft eine Nukleotidsequenz, wie sie in SEQ ID NO: 1, SEQ ID NO: 2 oder SEQ ID NO: 3 beschrieben ist. Zunächst wurde eine molekulare Epidemiologie-Studie unter mehr als 100 IDUs aus China, die seropositiv bezüglich des Subtyps

10 C von HIV-1 waren, durchgeführt, um notwendige Informationen zu sammeln über repräsentative virale Genome von im wesentlichen voller Länge. Die Genotypisierung auf der Basis der konstanten Region 2 und der variablen Region 3 (C2V3) innerhalb des Gens für das virale Hüll-Glykoprotein offenbarte die höchste Homologie der am meisten prävalenten Virusstämme, die in ganz China zirkulieren, zu Sequenzen des Subtyps C indischen Ursprungs.

15 Basierend auf diesen Ergebnissen wurde aus peripheren mononukleären Blutzellen (PBMC) von einem ausgewählten HIV-infizierten IDU direkt ein Genom von im wesentlichen voller Länge amplifiziert und subkloniert, das die am meisten prävalente Klasse der C-Stämme, die in ganz China zirkulieren, darstellt. Die Sequenzanalyse identifizierte eine Mosaikstruktur, was auf extensive Intersubtyp-Rekombinationsvorgänge zwischen den Genomen der prävalenten C- und

20 (B')-Subtyp-Thai-Virusstämme jener geographischen Region deutet. Eine RIP-Analyse (Recombinant Identification Program Analysis) und phylogenetisches "bootstrapping" legten insgesamt 10 Bruchstellen (i) in der für gagpol kodierenden Region, (ii) in vpr und am 3'-Ende des vpu-Gens und (iii) im offenen Leserahmen von nef nahe. Thai (B')-Sequenzen umfassen daher (i) mehrere Insertionen in der kodierenden Region von gagpol (Nukleotide 478-620, 1290-

25 1830, 2221-2520, jeweils bezogen auf das erste Nukleotid innerhalb des Startkodon des Gag- bzw. des GagPol-Leserahmens), (ii) 3'-vpr, das komplette vpu, die ersten Exons von tat und rev (etwa 1000 Nukleotide beginnend etwa an Nukleotid 138 bezogen auf das Startkodon des Vpr-Leserahmens) und (iii) die 5'-Hälfte des nef-Gens (Nukleotide 1-300). Die übrigen Bereiche innerhalb der 9078 Nukleotide umfassenden Sequenz (SEQ ID NO: 1; Tabelle 3) weisen höchste

30 Homologien zu bekannten Subtyp C Isolaten auf. Bruchstellen von 97cn54, die in der kodierenden Region von vpr/vpu bzw. im nef-Gen lokalisiert sind, wurden bei vielen Stämmen des Subtyps C, die von IDUs isoliert wurden, die in verschiedenen Gebieten Chinas leben, an

- ähnlichen Positionen gefunden. Dies legt eine gemeinsame Abstammung für die C/B'-rekombinanten Stämme nahe. Bei mehr als 50% der gut definierten CTL-Epitope, die vom Subtyp B abstammen, innerhalb von Gag und Pol und bei 10% der bekannten Epitope in Env, wurde gefunden, daß die Sequenzen innerhalb dieser C/B'-chimären Referenzstämme exakt
- 5 übereinstimmen. Diese Ergebnisse können die Anstrengungen in bezug auf Impfstoffe in China deutlich erleichtern, indem außerordentlich wichtige Matritzen für die Konzeption von Impfstoffen bereitgestellt und Reagenzien für die am besten geeigneten immunologischen/virologischen Ausleseverfahren entwickelt werden.
- 10 Die Verwendung der beschriebenen Sequenz gemäß vorliegender Erfindung, einer Sequenz von HIV-1, das den am meisten prävalenten C-Typ Virusstamm innerhalb Chinas darstellt, als Grundlage und Ausgangsmaterial ist für die Entwicklung von präventiv oder therapeutisch einsetzbaren Impfstoffen von Vorteil. Die notwendigen Konsequenzen für die Entwicklung eines erfolgreichen HIV-Impfstoffkandidaten sind (i) ein detailliertes Wissen über die jeweilige
- 15 epidemiologische Situation und (ii) die Verfügbarkeit einer klonierten kodierenden Sequenz, die innerhalb einer geographischen Region oder einer bestimmten Bevölkerung den am meisten prävalenten Virusstamm repräsentiert. Solche Sequenzen stellen die Grundlage dar (i) für die rationale Konzeption von präventiv und therapeutisch einsetzbaren HIV-Impfstoffkandidaten, (ii) für Entwicklung spezifischer Therapeutika, wie beispielsweise therapeutisch wirksamer Decoy-
- 20 Oligonukleotiden und Proteine, Antisense-Konstrukte, Ribozyme und transdominant negativ wirksamer Mutanten (iii) für die Entwicklung lentiviraler Vektoren für die Gentherapie und (iv) die Herstellung von Reagenzien, die für Diagnose und Verlaufskontrolle der HIV-Infektion sowie die immunologische/virale Überwachung des Impfungsprozesses eingesetzt werden können.
- 25 Dies ist insbesondere zutreffend für Impfstoffkandidaten, die auf den HIV-Hüllproteinen beruhen, von denen gezeigt wurde, daß sie unter allen HIV-Proteinen die größte Variabilität aufweisen. Darüber hinaus wird ein erfolgreicher Impfstoff sehr wahrscheinlich beide Arme des Immunsystems induzieren müssen: neutralisierende Antikörper, idealerweise gerichtet gegen
- 30 Konformations-Epitope im Hüllprotein sowie zellvermittelte Immunantworten (CD4-positive T-Helfer-Zellen, CD8-positive zytolytische T-Zellen, Zytokine vom Typ Th-1,  $\beta$ -Chemokine), erzeugt gegen Epitope verschiedener viralen Proteine. Das Konformations-Epitop gemäß der

- vorliegenden Erfindung besteht aus mindestens 3 Aminosäuren, vorzugsweise aus 5 oder mehr Aminosäuren, die bei der Antikörper-Bindung involviert sind. Konformationelle Epitope können sich auch aus mehreren Abschnitten entweder eines einzigen Proteins, oder - im Falle oligomerer Komplexe wie z.B. des trimeren Hüllglykoprotein-Komplexes - aus mehreren Abschnitten unterschiedlicher Untereinheiten zusammensetzen. Ein lineares Epitop gemäß der vorliegenden Erfindung variiert normalerweise in der Länge und umfaßt mindestens 8 Aminosäuren bis etwa 15 Aminosäuren oder mehr, wobei eine Länge von 9 bis 11 Aminosäuren insbesondere im Falle MHC Klasse I restringierter CTL-Epitope bevorzugt ist.
- 10 Die vorliegende Erfindung betrifft somit ferner Polypeptide, kodiert von der Nukleotidsequenz oder Fragment oder Derivat der Nukleotidsequenz gemäß SEQ ID NO:1, 2 oder 3. Die vorliegende Erfindung betrifft ferner Polypeptide, umfassend eine kontinuierliche Sequenz von mindestens 8 Aminosäuren, die von der Nukleotidsequenz oder Fragmenten oder Derivaten der Nukleotidsequenz gemäß SEQ ID NO:1, 2 oder 3 kodiert werden. Vorzugsweise umfasst das erfindungsgemäße Polypeptid eine antigene Determinante, die natürlicherweise in Infizierten eine Immunreaktion auslöst. Besonders bevorzugt sind Polypeptide, umfassend eine Aminosäuresequenz, kodiert von der Nukleotidsequenz gemäß SEQ ID NO:2 oder 3 oder dessen Derivate und Fragmente. Insbesondere bevorzugt sind Epitope umfassend einen kontinuierlichen Bereich von 9 bis 11 Aminosäuren, die identisch sind zwischen den durch SEQ ID NO:1 kodierten Polypeptiden und einem HIV-1<sub>LAI</sub> Referenzisolat, oder die 2 oder weniger konservierte Aminosäuresubstitutionen innerhalb der 9 bis 11 Aminosäuren umfassenden Sequenz aufweisen. Beispiele für derartige Epitope sind in Beispiel 11 aufgeführt. Die erfindungsgemäßen Polypeptide können z.B. als Impfstoffe und Therapeutika oder zur Diagnostik verwendet werden.
- 15 20 25 30
- Ein weiterer Aspekt der vorliegenden Erfindung betrifft ein Polynukleotid gemäß SEQ ID NO:1, 2 oder 3. Ferner betrifft die vorliegende Erfindung ein Polynukleotid-Fragment der Nukleotidsequenz gemäß SEQ ID NO:1, 2 oder 3, oder ein Polynukleotid, das mindestens eine kontinuierliche Sequenz von Nukleotiden umfaßt, die zur selektiven Hybridisierung an die Nukleotidsequenz, wie sie in SEQ ID NO:1, 2 oder 3 dargestellt ist, in der Lage ist. Ferner betrifft die vorliegende Erfindung Derivate der erfindungsgemäßen Polynukleotide oder Polynukleotid-Fragmente. Vorzugsweise umfasst das Polynukleotid oder das Polynukleotid-Fragment eine kontinuierliche Sequenz von mindestens 9 Nukleotiden, bevorzugterweise von

mindestens 15 Nukleotiden, noch bevorzugterweise von mindestens 27 Nukleotiden, oder eine längere Sequenz. Das Polynukleotid oder das Polynukleotid-Fragment kann auch die kodierende Region der einzelnen HIV-Gene umfassen, wie z.B. von gag, pol, env. Beispiele sind in SEQ ID NO: 2 und SEQ ID NO: 3 angegeben. Ein weiterer Aspekt der vorliegenden Erfindung betrifft 5 ein Polynukleotid, umfassend mindestens 2 erfindungsgemäße Polynukleotid-Fragmente, wobei die Sequenzen der Polynukleotid-Fragmente auch überlappen oder durch einen Nukleotid-Platzhalter voneinander getrennt sein können. Die Sequenzen der Polynukleotid-Fragmente können identisch oder verschieden sein. Die erfindungsgemäßen Polynukleotide oder Polynukleotid-Fragmente können als Impfstoffe oder Therapeutika oder zur Diagnostik 10 verwendet werden.

Die kodierende Sequenz des Klons 97cn54 und Derivate davon, ausgeführt in Form der SEQ ID NO: 1, als Vertreter des HIV-1 vom Subtyp C kann als Grundlage für die folgenden Anwendungen verwendet werden:

15 Entwicklung von Subtyp-C-spezifischen HIV-1-Impfstoffen für prophylaktische und therapeutische Zwecke. Diese Subtyp-spezifischen Impfstoffe können weltweit in allen geographischen Regionen, wo das Subtyp-C-Virus für die HIV-Epidemie eine wesentliche Rolle spielt, verwendet werden, also z.B. in Lateinamerika, in Afrika und in Asien. Insbesondere 20 sollten HIV-Impfstoffe, die getestet werden sollen in und entwickelt werden sollen für Südost-Asien und China auf der beschriebenen kodierenden Sequenz von 97cn54 beruhen, um Subtyp-spezifische humorale und zellvermittelte Immunantworten zu induzieren. Des Weiteren können solche HIV-1 Subtyp C-spezifischen Impfstoffe als eine Komponente in einer Kocktail-Vakzine eingesetzt werden, die entweder alle oder eine definierte Auswahl der weltweit relevanten HIV 25 Subtypen berücksichtigt.

Um gute humorale und zellvermittelte Immunantworten in den Impflingen zu induzieren, enthalten die Antigene oder kodierenden Sequenzen, die dem Immunsystem zugeführt werden sollen, vorzugsweise (i) kurze kontinuierliche Abschnitte von mindestens 3 bis etwa 5 30 Aminosäuren Länge oder längere Abschnitte, abgeleitet von einem der offenen Leserahmen, wie sie in Tabelle 3 abgebildet sind, (ii) Bereiche von vorzugsweise 9 bis 11 Aminosäuren, (iii) Kombinationen dieser Bereiche, die entweder getrennt oder als Polypeptid-Kette (*Epitope*-

- Strings)* verabreicht werden, wobei die Epitope-Strings bzw. deren Aminosäuresequenzen entweder überlappen oder aber durch Aminosäuren oder andere Platzhalter getrennt sein können, und besonders bevorzugterweise vollständige Proteine oder die entsprechenden kodierenden Sequenzen oder deren Varianten, die auch umfangreiche Deletionen umfassen können. Daher
- 5 betrifft eine andere Aufgabe der vorliegenden Erfindung Polypeptide, die kodiert werden von den Nukleotidsequenzen oder Fragmenten der Nukleotidsequenzen, wie sie in SEQ ID NO: 1, SEQ ID NO: 2 und SEQ ID NO: 3 dargestellt sind. Vorzugsweise umfaßt das Polypeptid eine kontinuierliche Sequenz von mindestens 8 Aminosäuren, vorzugsweise mindestens von 9 bis 11 Aminosäuren, besonders bevorzugterweise von mindestens 15 Aminosäuren oder längere
- 10 Sequenzen oder diskontinuierliche Epitope, die sich vorzugsweise aus wenigstens drei Aminosäuren einer einzigen Polypeptidkette oder, im Falle oligomerer Proteinkomplexe, auch unterschiedlicher Polypeptidketten zusammensetzen. Impfstoff-Konstrukte auf der Basis der kodierenden Sequenz von 97cn54 schließen alle im Stand der Technik bekannten Antigenformen ein und greifen auf einschlägige Verabreichungssysteme zurück.
- 15 Kurze Epitope, kodiert von Fragmenten der Nukleinsäuresequenzen gemäß SEQ ID NO: 1 bis 3, und jeweils drei bis fünf Aminosäuren, vorzugsweise von 9 bis 11 oder mehr Aminosäuren umfassend, können vorzugsweise synthetisch hergestellt werden. Derartige Peptide enthalten entweder ein B-Zellepitop, ein MHC Klasse II-restringiertes T-Helferepitop, ein MHC Klasse I-restringiertes zytotoxisches T-Zellepitop oder Kombinationen der genannten Varianten. Dabei können einzelne Epitope überlappen oder auch durch Platzhalter, präferentiell bestehend aus Glyzin und/oder Serin Resten voneinander getrennt sein. Verzweigtkettige Peptide können entsprechend dem Stand der Technik entweder während der Synthese oder unter Zuhilfenahme der gängigen und kommerziell erhältlichen homo- und heterobifunktionellen chemischen
- 20 Quervernetzer im Anschluß an die Synthese und Reinigung der entsprechenden Peptide erzeugt werden. Alternativ können *per se* wenig immunogene Peptide durch Quervernetzung auch an ausgewählte Trägerproteine wie z.B. Ovalbumin konjugiert werden, gentechnisch in Trägerproteine inseriert oder an deren N- bzw. C-Terminus fusioniert werden. Vorzugsweise sind derartige Trägerproteine (i) bei Expression in geeigneten Zellkultursystemen (siehe unten) oder
- 25 (ii) nach geeigneter Rückfaltung des gereinigten, denaturierten Proteins in der Lage, partikuläre Strukturen auszubilden, bei denen B-Zellepitope vorzugsweise auf der Oberfläche des partikulären Carriers zu liegen kommen. Zahlreiche Beispiele solcher zur Ausbildung

- partikulärer Strukturen tendierender Polypeptide sind mittlerweile bekannt wie beispielsweise das Hepatitis B-Virus (HBV) Core Antigen (HBcAg), das HBV Oberflächenprotein (HBsAg), das HIV gruppenspezifische Antigen, das Polyomavirus VP1 Protein, das Pappillomvirus L1 Protein oder das TyA Protein der Hefe. Aufgrund der Tatsache, daß sich die Mehrheit der bislang beschriebenen partikelbildenden Proteine aus den Kapsid- oder Strukturproteinen unterschiedlichster Viren rekrutiert, spricht man hier auch von Virus-ähnlichen Partikeln (VLP, virus-like particles; Übersicht: Sonderausgabe *Vaccine*. (1999) Volume 18. Advances in Peptide, Protein and Nucleic Acid Vaccine Strategies. edited by Prof. P.T.P. Kauyama)
- 10    *Epitop-Strings* und Polypeptide, kodiert von Fragmenten der Nukleinsäuresequenzen gemäß SEQ ID NO: 1 bis 3, mit einer Länge größer 30, vorzugsweise größer 50 Aminosäuren sowie Polypeptide mit einer Tendenz zur Ausbildung partikulärer Strukturen (VLP) können nach dem Stand der Technik in Prokaryonten produziert und gereinigt werden. Derartige Plasmide enthalten dementsprechend einen bakteriellen Replikationsursprung wie z.B. ColE1, in aller Regel einen Selektionsmarker wie z.B. eine Resistenz gegenüber Kanamycin oder Ampicillin, eine konstitutiv aktive oder induzierbare Transkriptions-Kontrolleinheit wie beispielsweise den LacZ- oder Tac Promotor, sowie die Signale zum Translationsstart und Translationsstop. Zur vereinfachten Expression und Affinitätsreinigung können auch optional abspaltbare Fusionsanteile und Reinigungshilfen wie beispielsweise die Glutathion-S-Transferase oder 15    Reinigungshilfen wie z.B. Oligohistidin-tags (Fänger) verwendet werden.
- 20

Die DNA- oder RNA-Sequenzen, die (i) zur Herstellung der *Epitop-Strings*, kompletter Proteine oder Virus-ähnlicher Strukturen in eukaryontischen Zellkulturen wie z.B. Hefezellen, Pilzen, Insektenzellen oder Säugerzellen verwendet werden oder die (ii) zur direkten Verabreichung von 25    DNA zu Immunisierungszwecken eingesetzt werden, können sich auf eine Verwendung der Kodons verlassen, wie sie vom Virus selbst verwendet wird. Alternativ kann die Verwendung der Kodons, wo immer technisch möglich, angepasst werden an die am häufigsten oder zweithäufigsten verwendeten Kodons in Genen, die im jeweiligen Produktionssystems hoch exprimiert werden. Beispiele für die Optimierung des Kodongebrauchs in einem unter 30    Sicherheitsaspekten optimierten Polygen, beinhaltend die Gene Gag, Pol und Nef, sowie im Hüllprotein-Gen sind gegeben in SEQ ID NO: 2 und SEQ ID NO: 3. Die SEQ IDs NO: 2 und 3 sind in Beispiel 15 näher spezifiziert.

Die Etablierung von Zelllinien zur Produktion der *Epitop-Strings*, Polypeptide oder Virusähnlichen Strukturen in den genannten Zellkultur-Systemen kann dem Stand der Technik entsprechend auf Vektoren basieren, die wiederum neben einem bakteriellen  
5 Replikationsursprung, einem positiven oder negativen Selektionsmarker vor allem die entsprechenden Kontrollregionen zur regelkonformen Transkription und Translation des Fremdproteins beinhalten können. Die nachfolgend beschriebenen Komponenten der DNA  
Vakzinkonstrukte stehen exemplarisch auch für die Module, die sich auch in Vektoren zur Expression der *Epitop-Strings*, Polypeptide oder kompletten Proteine in unterschiedlichen  
10 Säugerzellkulturen wiederfinden.

Bei der einfachsten Form der Immunisierung handelt es sich um die direkte Verabreichung eines reinen DNA-Impfstoffes. Dieser enthält im wesentlichen 5'-seitig vom kodierenden Bereich eine Transkriptions-Kontrollregion, auch Promotor/Enhancer-Region genannt, der wahlweise ein  
15 funktionelles Intron zur Steigerung der Genexpression folgen kann, (ii) eine Kozak-Sequenz inklusive eines Translations-Startkodons sowie am 3'-Ende des Fremdgens ein Translations-Terminationskodon gefolgt von einer Polyadenylierungs-Signalsequenz. Die Promotor/Enhancerregion kann präferentiell eine konstitutive Expression des gewünschten Genproduktes unterstützen und ist beispielsweise von der Transkriptions-Kontrollregion eines  
20 unmittelbar frühen (IE) Cytomegalievirus-Gens (CMV-IE) oder dem Rous-Sarcoma Virus (RSV) LTR (*long terminal repeat*) abgeleitet. Alternativ kann auch eine induzierbare Form einer Transkriptions-Kontrollregion wie z.B. ein *Tet on/Tet off* Promotor verwendet werden, bei dem die Transkription beispielsweise durch die Gabe von Tetrazyklin oder entsprechender Analoga reguliert wird. Des Weiteren bietet sich hier die Verwendung von Zelltyp-spezifisch regulierten  
25 Transkriptions-Kontrollregionen an wie z.B. die stromaufwärts des Muskel-Kreatin-Kinase Gens (MCK Gen; muskelspezifische Expression), des CD4-Rezeptorgens oder der MHC Klasse II Gene (präferentielle Expression in Antigen-präsentierenden Zellen) gelegenen Promotor/Enhancerregionen. In einigen Fällen verwendet man auch chimäre Kombinationen aus  
30 (i) Zelltyp-spezifischen Promotoren und (ii) viralen Enhancerregionen, um die Vorteile einer gewebespezifischen Expression mit denen der starken Transkriptionsaktivität viraler Enhancer zu vereinen. Die Verstärkung der Genexpression durch das Einbinden eines in aller Regel 5'-seitig des offenen Leserahmens gelegenen funktionellen Introns geht auf eine gesteigerte

Kernexportrate gespleißter im Vergleich zu ungespleißten Transkripten zurück und wird beispielsweise durch die Insertion eines im  $\beta$ -Globin Gen gelegenen Introns erreicht.

- Eine bevorzugte Form eines auf SEQ ID NO:1, 2 oder 3 basierenden DNA-Impfstoffs enthält
- 5 zusätzlich ein von Alpha-Viren wie beispielsweise von Semliki-Forest- oder Venezuela-Encephalitis-Viren (SFV, VEE) abgeleitetes Replikon. In diesem Fall folgt der oben beschriebenen nukleären Transkriptions-Kontrolleinheit und dem wahlweise berücksichtigten Intron zunächst der für die VEE oder SFV Nichtstrukturproteine (NS) kodierende Bereich. Erst
- 10 3' seitig davon folgt das eigentliche Fremdgen, dessen zytoplasmatische Transkription seinerseits durch einen NS-sensitiven Promotor reguliert wird. Dementsprechend wird ausgehend von der nukleären Transkriptions-Kontrolleinheit ein langes Transkript über mehrere offene Leserahmen erzeugt, das anschließend ins Zytosol transloziert wird. Die dort synthetisierten NS-Proteine aktivieren dann durch Bindung an die entsprechende Kontrollregion die zytoplasmatische Transkription der Fremdgene. Dieser Amplifikationseffekt führt in aller Regel zu einer
- 15 abundanten RNA Synthese und folglich hohen Fremdprotein-Syntheseraten. Letzteres erlaubt, im direkten Vergleich mit konventionellen Plasmiden, die auf den beschriebenen Effekt durch zytoplasmatische RNA Amplifikation verzichten, in aller Regel eine deutliche Reduktion der zu verabreichen Plasmidmenge bei wenigstens vergleichbarer Immunogenität.
- 20 Die oben beschriebenen Peptide, Proteine, Virus-ähnlichen Partikel und DNA-Konstrukte können durch intramuskuläre, subkutane, intradermale, intravenöse Injektion verabreicht werden, wobei für die Verabreichung der proteinösen Antigene jeweils der Stand der Technik angewendet wird. Zur DNA-Immunisierung können entweder konventionelle Spritzen mit Injektionsnadeln verwendet werden, oder aber Gerätschaften, die ohne Nadeln auskommen und in aller Regel die
- 25 DNA über Druckluft direkt in das gewünschte Gewebe einbringen können. Dazu zählt insbesondere auch die intranasale und orale Applikation DNA-haltiger Vakzin- Formulierungen durch sprayartige Vorrichtungen. Alternativ dazu kann die DNA auch an feste Träger wie z.B. Goldkügelchen konjugiert und beispielsweise unter Luftdruck in die entsprechenden Gewebe verabreicht werden.
- 30 Zur Verstärkung oder Modulation der Immunantwort können die erwähnten proteinösen Antigene und DNA-Konstrukte auch mit sogenannten Adjuvanten, i.d. Regel Stimulatoren der

Immunantwort, kombiniert oder in einer sequentiellen Abfolge mit den Adjuvantien verabreicht werden. Konventionelle Adjuvantien wie z.B. Aluminiumhydroxyd oder Aluminium-Hydroxyphosphat resultieren in einer Stimulation der humoralen Immunantwort, die sich auch durch hohe Antikörpertiter vom IgG1 Subtyp auszeichnet. Modernere Adjuvantien, wie 5 beispielsweise CpG Oligonukleotide (Konsensuskernmotiv: Purin-Purin-CpG-Pyrimidin-Pyrimidin) oder chemisch modifizierte Derivate davon (Phosphothioat-Oligonukleotide; Oligonukleotide mit Peptidrückgrat) verstärken üblicherweise den zellulären Arm der Immunantwort und unterstützen vornehmlich den Th1-Typ der zellvermittelten Immunität, gekennzeichnet durch hohe Antikörpertiter vom Subtyp IgG2a und die Induktion von Th1 10 Zytokinen wie z.B.  $\gamma$ -IFN, IL-2 und IL-12.

Die Verabreichung und Aufnahme von Peptiden, Proteinen und DNA-Vakzinkonstrukten kann insbesondere auch verbessert werden durch Bindung an oder Inkorporation in höhermolekulare Strukturen, wie z.B. biodegradierbare Partikel, multilamellare, idealerweise kationische 15 Liposomen, immunstimulierende Komplexe (ISCOMS), Virosomen oder *in vitro* assemblierter Viruspartikel. Zu biodegradierbaren Partikeln zählen beispielsweise PLA- (L-lactic acid), PGA- (polyglycolic) oder PLGA- [poly(D,L-lactide-co-glycolide)] Mikrosphären oder Derivate davon, kationische Mikropartikel oder von bakteriellen Kapselpolysacchariden abgeleitete Trägersubstanzen. Der Sammelbegriff ISCOMS steht für immunstimulierende Komplexe, die auf 20 wasserlöslichen Extrakten der Rinde von *Quillaja saponaria* entstammen und mittels chromatographischer Verfahren weiter aufgereinigt wurden. Eine dem Stand der Technik entsprechende, detaillierte Übersicht zu den unterschiedlichsten Adjuvantien und Verabreichungshilfen findet sich unter 25 <http://www.niaid.nih.gov/aidsvaccine/pdf/compendium.pdf> [Vogel, F. R., Powell, M. F. and Alving, C. R. „A Compendium of Vaccine Adjuvants and Excipients (2nd Edition)].

Des Weiteren können zur günstigen Präsentation von Epitop-Strings, Polypeptiden und Virusähnlichen Partikeln virale und, alternativ, bakterielle Vektoren eingesetzt werden.

30 Nach dem aktuellen Stand der Technik eignen sich beispielsweise gentechnisch veränderte Salmonellen und Listerien aufgrund ihres natürlichen Zelltropismus in besonderer Weise dazu, DNA-Vakzinkonstrukte in Antigen-präsentierende Zellen wie Monozyten, Makrophagen und

- vor allem in dendritische Zellen einzubringen. Die gentechnischen Veränderungen können neben einem Gewinn an Zelltypspezifität unter anderem dazu beitragen, daß die DNA unbeschadet das Zytoplasma der Antigen-präsentierenden Zelle erreicht. In diesem Fall gelangt ein DNA-Vakzinkonstrukt in den Zellkern, wo über einen eukaryontischen, vorzugsweise viralen oder zelltypspezifischen Promotor der entsprechende Leserahmen unter Nutzung der zellulären Resourcen und Proteine transkribiert wird. Nach dem Transport der RNA ins Zytoplasma wird das entsprechende Genprodukt translatiert und, je nach Beschaffenheit, posttranslational modifiziert und dem entsprechenden zellulären Kompartiment zugewiesen.
- 10 Bakterielle Vektoren (Salmonellen, Listerien, Yersinien etc.) können auch zur Induktion einer Schleimhautimmunität, vorzugsweise nach oraler Verabreichung verwendet werden. Dabei werden die entsprechenden Antigene durch die bakterielle Transkriptions- und Translationsmaschinerie hergestellt und unterliegen demnach nicht den in Säugerzellen sonst üblichen posttranslationalen Modifikationen (keine entsprechende Glykosylierung; kein sekretorischer *Pathway*).

Daneben existieren mittlerweile eine Vielzahl von attenuierten viralen Vektoren, mit deren Hilfe sich die gewünschten Antigene erfolgreich und in hohen Ausbeuten exprimieren lassen. Neben deren Tauglichkeit zur reinen Antigen-Produktion können solche virale Vektoren auch direkt zur Immunisierung eingesetzt werden. Diese kann zunächst entweder *ex vivo* erfolgen, beispielsweise zur Infektion von Antigen-präsentierenden Zellen, die anschließen dem Impfling verabreicht werden, oder direkt *in vivo* durch die subkutane, intradermale, intracutane, intramuskuläre oder intranasale Immunisierung mit dem rekombinanten Virus, die eine günstige Antigen-Präsentation mit entsprechendem Immunisierungserfolg erzielen lässt. So können beispielsweise durch Immunisierung mit rekombinannten Vakzinien Viren wie z.B. dem durch Passagieren über Hühnerzellen attenuierten *Modifizierten Vaccinia Virus Ankara* (MVA), dem gentechnisch attenuierten *Vaccinia Stamm New York* (NYVAC) oder die in Vögeln endemischen aviären *Vaccinia* Viren (*Fowlpox, Canarypox*) adäquate humorale und zellvermittelte Immunantworten in den geimpften Personen induziert werden. Alternativ eignen sich dazu in gleicher Weise auch eine Reihe anderer Viren wie z.B. rekombinante Alpha-Viren, darunter das Semliki-Forest Virus oder das Venezuela-Enzephalitis Virus, rekombinante Adenoviren, rekombinante Herpes Simplex Viren, Influenzaviren und andere.

Letztlich können auf Basis der SEQ ID NO:1, 2, oder 3 auch attenuierte HIV-Viren generiert und zu Immunisierungszwecken eingesetzt werden, sofern mittels Klonierverfahren nach dem Stand der Technik die Regulationssequenzen (LTR, long terminal repeat), die den kodierenden Bereich flankieren, ergänzt werden. Eine hinreichende Attenuierung des Virus kann dann entsprechend 5 dem Stand der Technik durch eine oder mehrere Deletionen beispielsweise im Nef-Gen erzielt werden.

Die in den Beispielen SEQ ID NO: 1 und SEQ ID NO: 3 ausgeführten Nukleinsäuresequenzen sowie daraus abgeleitete Peptide, Proteine oder Virus-ähnlichen Partikel können auch als 10 Komponenten viraler Vektoren zur Genüberführung eingesetzt werden.

Die durch das GagPol-Gen (SEQ ID NO:1; Nukleotid 177-4458; Tabelle 3) kodierten Polypeptide können beispielsweise die Verpackungs- und Rezeptorfunktionen von z.B. lenti- oder retroviralen Vektoren bereitstellen. So können z.B. nach transienter Transfektion von 15 Säugerzellen durch geeignete Plasmidvektoren, welche die gleichzeitige Expression des GagPol und VSV-G (vesicular stomatitis virus Hüllprotein G) Gens unterstützen und die Verpackung eines therapeutischen Transgens sicherstellen, Viruspartikel erzeugt werden, die auch in der Lage sind, ruhende, postmitotische oder enddifferenzierte Zellen zu transduzieren. Dieses Verfahren zur Generierung transduktionskompetenter Viruspartikel kann wesentlich erleichtert und 20 effizienter gestaltet werden, beispielsweise durch die Etablierung stabiler Zelllinien, z.B. basierend auf *human embryonic kidney* Zellen (HEK293), die das GagPol Polyprotein konstitutiv oder unter Kontrolle eines induzierbaren Promoters exprimieren. Alternativ können auch rekombinante Adenoviren generiert werden, die die Verpackungsfunktionen, die Rezeptorfunktionen und die Transgenfunktionen oder Kombinationen daraus kodieren, und so 25 als Werkzeug zum *ex vivo*, *in situ* und *in vivo* Delivery von retro- oder lentiviralen Vektoren dienen.

Die durch SEQ ID NO: 3 kodierten Hüllproteine oder Derivate davon können die Rezeptorfunktion für lenti-, spuma- oder retrovirale Vektoren oder anderer, auf umhüllten Viren 30 basierender Vektoren durch Inkorporation in den Lipid-Bilayer bereitstellen. Dazu können beispielsweise auch Verpackungslinien erzeugt werden, bei denen sowohl die GagPol Proteine von Retro-, Spuma- und vorzugsweise von Lentiviren, als auch die aus SEQ ID NO: 1 und 3

abgeleiteten Hüllproteine entweder konstitutiv oder unter Kontrolle eines induzierbaren, wahlweise eines in der Aktivität regulierbaren Promoters exprimiert werden. Alternativ dazu können, beispielsweise basierend auf dem Genom von Typ C oder Typ D Retroviren oder anderer membranumhüllter Viren wie z.B. Influenza- oder Herpesviren, chimäre Viren generiert werden, die zusätzlich zu dem natürlichen Hüllprotein oder anstelle des natürlichen Hüllproteins ein von SEQ ID NO: 1 oder SEQ ID NO: 3 abgeleitetes Hüllprotein auf der Oberfläche tragen.

Gegen die aus den SEQ IDs NO: 1 bis 3 abgeleiteten Peptide, Proteine oder Virus-ähnlichen Partikel können auch (i) polyklonale Antiseren, (ii) monoklonale Antikörper (Maus, Mensch, 10 Kamel), (iii) Antikörperderivate wie beispielsweise *single-chain* Antikörper, humanisierte Antikörper, bispezifische Antikörper, Phagen-Antikörperbanken oder (iv) andere hochaffin bindende Polypeptide wie z.B. Derivate des hPSTI (human pancreatic secretory trypsin inhibitor) generiert werden. Diese Reagentien können zu therapeutischen Zwecken, beispielsweise zur Behandlung von HIV-Infektionen oder zu diagnostischen Zwecken, beispielsweise zur 15 Herstellung von Testkits verwendet werden.

Auf ähnliche Weise können die aus SEQ ID NO:1, 2 oder 3 abgeleiteten Peptide, Proteine oder Nukleinsäure-Sequenzen für diagnostische Zwecke, z.B. für die Serodiagnostik und für die Anwendung von Nukleinsäure-Hybridisierungstechniken oder Nukleinsäure- 20 Amplifikationssystemen oder Kombinationen davon verwendet werden. Vorzugsweise können die erfindungsgemäßen Polynukleotid-Fragmente der Nukleotidsequenz gemäß SEQ ID NO:1 in einer Polymerase-Kettenreaktion eingesetzt werden. Besonders bevorzugt werden die erfindungsgemäßen Polynukleotid-Fragmente der Nukleotidsequenz gemäß SEQ ID NO:1 zur Diagnostik mittels DNA-Chiptechnologie eingesetzt.

25 Die Erfindung wird durch die nun folgenden Beispiele erläutert, ist aber nicht auf diese beschränkt:

**Beispiel 1:**

30 Blutproben  
Alle Blutproben, die für diese Studie verwendet wurden, wurden im Zuge der nationalen, molekularepidemiologischen Studie von 1996/1997 bezüglich HIV-1, Subtyp C, seropositiven

IDUs aus mehreren HIV-epidemischen Gebieten in China entnommen. Periphere mononukleäre Blutzellen (PBMC) wurden mittels Ficoll-Gradienten abgetrennt. Die Viren wurden durch Kokultivierung der PBMCs von seropositiven IDUs mit Phytohämagglutinin (PHA)-stimulierten Donor-PBMCs isoliert. Positive Viruskulturen wurden aus den Zellkultur-Überständen 5 nachgewiesen mittels des HIV-1 p24 Core Profile ELISA-Kits (DuPont Inc., Boston, MA).

**Beispiel 2:****Polymerase Kettenreaktionen (PCR) und DNA-Sequenzierung**

Provirale DNA wurde aus produktiv infizierten PBMCs von mehr als einhundert ausgewählten 10 HIV-1-positiven IDUs aus den nordwestlichen Provinzen Chinas extrahiert (Qiagen Inc., Valencia, CA). Die *Nested*-PCR wurde verwendet, um die kodierende Region für env C2V3 zu amplifizieren. Die PCR-Produkte wurden mittels der *Taq-cycle*-Methode unter Verwendung von Fluoreszenzfarbstoff-markierten Terminatoren (Applied Biosystems, 373A, Foster City, CA) wie kürzlich beschrieben (Bai et al. 1997; Yu et al. 1997) direkt sequenziert. Multiple 15 Sequenzvergleiche wurden unter Verwendung der *Wisconsin software package Genetics Computer Group* mit den Korrekturmethoden nach Kimura durchgeführt (GCG, 1997, Version 9).

**Beispiel 3:**

20 Phylogenetische Stammbaum-Analysen wurden von allen erhaltenen Sequenzen unter Verwendung des PHYLIP-Software-Pakets durchgeführt. Evolutionäre Entfernung wurden an Hand der *maximum parsimony*-Methode berechnet und durch kumulative horizontale Länge der Zweige angegeben. Die statistische Robustheit des *neighbour joining* Stammbaums wurde wie kürzlich beschrieben durch *bootstrap resampling* überprüft (Graf et al. 1998).

25

**Beispiel 4:****Auswahl eines repräsentativen HIV-1-Isolats des Subtyps C von chinesischen IDUs**

Innerhalb der Gruppen betragen die berechneten durchschnittlichen Abstände innerhalb der für 30 C2V3 kodierenden Region auf DNA-Ebene  $2,26 \pm 1,43$ , was darauf hindeutet, daß die Epidemie in diesem Gebiet noch sehr jung ist. Die Unterschiede zwischen den Gruppen zwischen chinesischen Subtyp-C-Sequenzen und denen aus Indien, Afrika, und Südamerika betragen  $9,67 \pm 2,31$  (Indien),  $15,02 \pm 4,13$  (Afrika) und  $8,78 \pm 3,41$  (Südamerika). Das zeigt eine enge

phylogenetische Verwandtschaft zwischen indischen und chinesischen Subtyp-C-Sequenzen (Lole et. al. 1999) und eine nennenswerte genetische Entfernung zu der *per se* relativ heterogenen Gruppe afrikanischer HIV-1-Stämme des Subtyps C.

## 5 Beispiel 5:

Identifizierung eines Virus-Isolats, das den in China zirkulierenden prävalenten Virus-Stamm des Subtyps C am besten repräsentiert

Aus den analysierten Proben wurde ein als 97cn54 bezeichnetes repräsentatives Isolat identifiziert, das höchste Homologie (99,6%) zu einer berechneten Konsensus-Sequenz (cn-conV3), die auf Grundlage der charakterisierten lokalen HIV-Sequenzen (Tabelle 1) erstellt worden ist, aufweist. Multiple Aminosäure-Sequenzvergleiche einschließlich der Primärsequenzen der V3-Schleife von primären Subtyp-C-Vertretern aus den verschiedensten epidemischen Regionen und auch Konsensus-Sequenzen von anderen Subtypen (A-H, O, CPZ) unterstrichen den Subtyp-C-Charakter des ausgewählten Primärisolats 97cn54 (Tabelle 1). Verglichen mit einer V3-Gesamtkonsensus-Sequenz (consensus) zeigen sowohl 97cn54 als auch cn-con-c Aminosäure-Abweichungen an den Positionen 13 ( $H \rightarrow R$ ) und 19 ( $A \rightarrow T$ ), die beide charakteristisch für Isolate des Subtyps C sind (C consensus).

**Tabelle 1:** Aminosäure-Sequenz-Vergleich der V3-Schleifen:

	Position	1	11	21	31	38
	Consensus	CTRPNNNTRK	SIHIGPGQAF	YA---TGDII	GDIRQAH	C
5	C_94IN11246	-----	-r-----t-	--	--e-v	-n----
	C_93IN905	-----	-r-----t-	--	-----m	-----
	C_93IN999	-vr-----e	-r-----t-	--	--e--	-----
	C_consensus	-----	-r-----t-	-----	-----	-----
10	C_ind8	-----	-tr----t-	-----	-----	-----
	97cn54-v3	-----g	-----r-----t-	-----	-----	-----
	cn-con-v3	-----g	-----r-----t-	-----	-----	-----
	C_bro025	-----	-r-----	-----e	-----	-----
15	C_ind1024	-----	-r-----t-	-----	-----r-y-	-----
	C_nof	-----	r-rv-----tv	-----na	-----	-----
	C_zam20	-a-g-----	-r-----t-	f-----a	-----	-----
	C_sm145	---ya-----	-vr-----t-	-----n	-----	-----
	A_consensus	-----	-vr-----	-----	-----	-----
20	B_consensus	-----	-----r-	-t-----e	-----	-----
	D_consensus	---y-----q	rt-----l	-----tr-----	-----	-----
	E_consensus	---s-----t	-t-----v	-r-----	-----k-y	-----
	F_consensus	-----	-t-----	-----	-----k	-----
	G_consensus	-----	-t-----	-----	-----	-----
25	H_consensus	-----	-s-----	-----	-----k-y	-----
	O_consensus	-e-gidige	.-r---.m-w	-smglg-tng	nss-a-y	-----

**Tabelle 1:** Der Aminosäure-Sequenz-Vergleich der V3-Schleifen von Konsensus-Sequenzen verschiedener Subtypen von HIV-1 (A-O) und ausgewählte Isolate des Subtyps C aus verschiedenen Ländern. Die V3-Gesamtkonsensus-Sequenz wurde durch den Vergleich der Konsensus-Sequenzen von verschiedenen Subtypen (A-O) ermittelt. cn-con-V3 stellt die Konsensus-Sequenz von HIV-1-Stämmen Subtyps C, die in China prävalent sind, dar. 97cn54 wurde als repräsentatives Standard-Isolat der in China vorkommenden prävalenten HIV-1-Stämme des Subtyps C ausgewählt. "-" bedeutet keinen Austausch gegenüber der V3-Konsensus-Sequenz, Kleinbuchstaben bedeuten eine Aminosäure-Substitution und "." bedeutet Lücken. Alle Konsensus- und Isolat-Sequenzen für multiple Vergleiche wurden von der Datenbank Los Alamos erhalten.

**Beispiel 6:**

- 40 Die für das 97cn54 Hüllprotein kodierende Sequenz ist am nächsten verwandt mit Virus-Stämmen des Typs C aus Indien.  
 Phylogenetische Stammbaum-Analysen, ursprünglich basierend auf den C2V3-Sequenzen des

env-Gens, ergaben, daß sowohl 97cn54 als auch die Konsensus-Sequenz der chinesischen Isolate des Subtyps C sich mit den Stämmen des Subtyps C aus Indien (ind8, d1024, c-93in905, c-93in999, c-93in11246), aus Afrika (c-eth2220, c-ug286a2), und aus Südamerika (92br025, nof, cam20 und sm145) gruppieren (clustern). Dies weist darauf hin, daß die indischen Virus-Stämme des Subtyps C der Ursprung der Epidemie von HIV-1, Subtyp C, in China sein könnten (Figur 1). Diese Hypothese stimmt auch überein mit unserer früheren epidemiologischen Erkenntnis, die bestätigt, daß mit HIV-1, Subtyp C, infizierte Menschen in Yunnan Injektionskanülen mit indischen Schmuckhändlern im Grenzgebiet geteilt haben sollen (Shao et al. 1999).

10 **Beispiel 7:**

Klonierung des HIV-1-Genoms von im wesentlichen voller Länge

Genome von HIV-1 von im wesentlichen voller Länge wurden amplifiziert mittels des *Expand Long Template PCR*-Systems (Boehringer-Mannheim, Mannheim, Deutschland), wie beschrieben bei Graf et al. (1998) und Salminen et al. (1995). Die Startermoleküle (Primer) 15 wurden in konservierten Regionen innerhalb der langen terminalen Wiederholungen (LTR) von HIV-1 positioniert: TBS-A1 (5'-ATC TCT AGC AGT GGC GGC CGA A) und NP-6 (5'-GCA CTC AAG GCA AGC TTT ATT G). Gereinigte PCR-Fragmente wurden mit glatten Enden in einen mit *SrfI* verdauten pCR-Script-Vektor (Stratagene, Heidelberg, Deutschland) ligiert und in den *E. coli*-Stamm DH5 $\alpha$  transformiert. Verschiedene rekombinante Klone, die im wesentlichen 20 das HIV-1-Genom voller Länge enthielten, wurden mittels Restriktionsfragmentlängen-Polymorphismus (RFLP) und Sequenzierung der kodierenden Sequenz der V3-Schleife identifiziert. Laut RFLP-Analyse unter Verwendung verschiedener Kombinationen von Restriktionsendonukleasen und nachfolgender Sequenzierung der kodierenden Sequenz der V3-Schleife waren 77% der positiven Konstrukte voller Länge nahezu identisch. Ein Proviruss- 25 Konstrukt, das die breite Mehrheit der positiven Klone repräsentiert, wurde ausgewählt und wie oben beschrieben unter Verwendung des *primer-walking*-Ansatzes sequenziert (die Startermoleküle wurden ungefähr alle 300 bp entlang des Genoms für beide Strände entworfen).

**Beispiel 8:**

30 DNA-Sequenzen wurden unter Verwendung der Lasergene Software (DNASTAR, Inc., Madison, WI) auf Macintosh-Computern zusammengesetzt. Alle Referenzsequenzen der Subtypen dieser Studie sind von der Los Alamos HIV Datenbank. Ähnlichkeiten in der Nukleotid-Sequenz

wurden mittels des lokalen Homologie-Algorithmus von Smith und Waterman berechnet. Multiple Sequenzvergleiche mit verfügbaren Sequenzdaten anderer Subtypen wurden unter Verwendung des Wisconsin Softwarepakets *Genetics Computer Group* (GCG, 1997, Version 9) durchgeführt.

5

**Beispiel 9:****Gesamtstruktur der kodierenden Sequenz von 97cn54**

Die 9078 bp lange genomische Sequenz des Isolats 97cn54 enthielt alle bekannten strukturellen und regulatorischen Gene des HIV-1-Genoms. Es wurden keine wesentlichen Deletionen, 10 Insertionen oder Umlagerungen gefunden. Die Ähnlichkeiten in der Nukleotid-Sequenz wurden untersucht mittels Vergleich aller kodierenden Sequenzen (CDS) von 97cn54 mit Konsensus-Sequenzen verschiedener Genotypen und ausgewählter Subtyp-Isolate (Tabelle 2). Die höchsten Homologien der Leserahmen von gag, pol, env und vif zu den entsprechenden Konsensus-Sequenzen des Subtyps C lagen in einem Bereich von 93,93 bis 95,06%. Diese Beobachtung 15 erweiterte den oben beschriebenen Sequenzvergleich und die phylogenetische Stammbaum-Analyse aufgrund von C2V3 erheblich (siehe Tabelle 1 und Figur 1). Sie bestätigte daher eindeutig, daß das ausgewählte Virus-Isolat zur Gruppe der kürzlich veröffentlichten Virus-Stämme des Subtyps C gehört. Die durch diese Art der Analyse für die Gene tat, vpu, vpr und nef bestimmten Werte der Homologie waren jedoch nicht ausreichend, um eine klare Zuordnung 20 dieser Leserahmen zu Virus-Stämmen des Subtyps B oder C zu erlauben (Tabelle 2). Für das Gen vpu wurden die höchsten Homologien zu den Subtypen B registriert (94,24%), während die Homologie zu der Konsensus-Sequenz des Subtyps C nur 78,23% betrug. Ähnliche Beobachtungen wurden für das Gen tat gemacht: höchste Homologie zum Isolat B'-rl42 (>91%), im Vergleich zu 87,9% (C-92br025) und 85,5% (C-eth2220) für ausgewählte primäre Vertreter 25 des Subtyps C oder 89,01% für die Konsensus-Sequenz des Subtyps C. Diese Daten legten zusammen mit dem Auftreten der Genotypen B, C und E im ganzen epidemischen Gebiet von Yunnan nahe, daß das analysierte Virus-Isolat einen Mosaik-Virusstamm darstellen könnte, der die Folge eines Rekombinationsvorgangs zwischen Subtyp B' und Subtyp C ist.

30

**Tabelle 2:** Vergleich der kodierenden Sequenzen von 97cn54 mit den entsprechenden Genen von Referenz-Stämmen und Subtyp-spezifischen Konsensus-Sequenzen.

CDS	Prozent Identität mit 97cn54								
	gag	pol	vif	vpr	tat	rev	vpu	env	nef
A	87.68	91.80	86.81	83.66	84.90	83.97	79.82	85.75	84.19
B	90.43	91.93	88.04	90.31	86.56	82.08	94.24	84.52	88.13
B-mn	89.38	90.82	86.01	89.31	87.44	79.48	88.21	82.33	85.41
B'-rl42	91.53	90.76	86.01	88.97	91.163	80.23	96.74	82.70	85.99
C	94.65	94.29	95.06	91.39	89.01	91.99	78.23	93.93	88.82
C-92br025	92.19	92.91	88.51	90.03	87.91	89.70	76.13	88.51	86.20
C-eth2220	91.4	92.06	87.15	90.77	85.57	88.08	80.09	87.15	87.08
D	89.80	91.08	87.74	87.94	83.93	84.39	87.30	85.26	86.88
E/A	86.324	89.07	86.59	83.39	81.44	81.74	77.31	82.09	84.18
F	88.02	88.99	86.36	86.25	80.65	86.25	82.33	84.02	/
G	88.08	/	/	/	/	/	/	84.55	/
H	87.69	89.45	86.01	85.22	/	/	/	83.74	/
O	73.42	78.02	72.12	76.604	72.31	76.60	59.54	67.01	80.35
CPZ	74.14	78.80	93.75	75.44	76.00	75.44	64.41	72.42	/

- 5   **Tabelle 2:** Nukleotidsequenz-Vergleich aller kodierenden Sequenzen (CDS) zwischen 97cn54 und DNA-Sequenzen, die entweder (1) Konsensus-Sequenzen bestimmter HIV-1-Subtypen (erhalten von der Los Alamos HIV-Datenbank) oder (2) Isolate des Standard-Subtyps C (92br025 und eth2220) und B (mn und rl42) darstellen. Die Daten geben die Identität einer bestimmten Sequenz mit 97cn54 in Prozent an. Nicht-eindeutige Nukleotid-Positionen innerhalb der
- 10   Konsensus-Sequenzen wurden als identisch bewertet. Die höchsten Homologien sind in Fettdruck hervorgehoben. "/" bedeutet, daß von der Los Alamos Datenbank keine Konsensus-Sequenz verfügbar war.

**Beispiel 10:**

- 15   Bestimmung der Rekombinationen zwischen den Subtypen  
 Das rekombinante Identifikationsprogramm (RIP, Version 1.3; <http://hiv-wew.lanl.gov/tools>) wurde verwendet, um potentielle Mosaik-Strukturen innerhalb der Gesamtsequenz dieses Klons zu identifizieren (Fenstergröße: 200; Schwellenwert für die statistische Signifikanz: 90%; Umgang mit Lücken: STRIP; informativer Modus: OFF). Es wurden Lücken eingeführt, um den

Vergleich zu ermöglichen. Die Hintergrund-Sequenzen der Subtypen in dieser Analyse waren: u455 (Subtyp A), RL42 (chinesischer Subtyp B-Thai (B')), eth2220 (Subtyp C), z2d2 (Subtyp D), 93th2 (Subtyp A/E).

5   **Beispiel 11:**

Rekombination zwischen den Subtypen in der kodierenden Region für Gag-Pol von 97cn54

Auch wenn wesentliche Homologien zu den Virusstämmen des Subtyps C innerhalb der hochkonservierten Leserahmen von gag und pol beobachtet wurden, identifizierte die RIP-Analyse 3 Bereiche der intra-subtypischen Rekombination innerhalb gagpol um die Positionen

- 10   478-620, 1290-1830 und 2221-2520 oberhalb des Startkodons von gag. Diese verstreuten Abschnitte liegen innerhalb der Leserahmen von gag und pol und weisen höchste Homologien zu dem Prototyp B (Daten nicht gezeigt) und insbesondere zu einem Isolat des Subtyps B(B'), das aus Yunnan kommt (Figur 2), auf. Diese Beobachtung unterstreicht eindeutig die Wichtigkeit von RIP-Analysen, da einfache Homologie-Vergleiche auf der Basis von kompletten Genen nicht  
15   in der Lage waren, diese kleinen verstreuten Fragmente eines anderen Subtyps zu identifizieren. Um die mittels RIP-Analyse erhaltenen Daten zu bestätigen, erstellten wir mehrere phylogenetische Stammbäume unter Verwendung der Regionen, die die Bereiche der vorgeschlagenen Rekombination entweder flankieren oder überspannen (Figur 3). Unter Verwendung mehrerer Standard-Vertreter verschiedener Subtypen und einiger ausgewählter  
20   Primär-Isolaten des Subtyps C konnten alle vorgeschlagenen Bereiche der Rekombination bestätigt werden durch differenzielles Clustern von 97cn54 mit den jeweiligen Referenz-Isolaten der Subtypen C (Figuren 3 A, C, E, G) oder B (Figuren 3 B, D, F).

**Beispiel 12:**

- 25   Intersubtyp-Rekombination in der für env kodierenden Region von 97cn54

Wie die in Tabelle 2 zusammengefassten Sequenzvergleiche erwarten ließen, bestätigte die RIP-Analyse die Intersubtyp-Rekombination zwischen Subtyp (B')-Thai und C (Figur 4) eindeutig. Ein Fragment von etwa 1000 bp Länge, das sich von den 150 3'-terminalen bp von vpr über das erste Exon von tat und rev bis zu vpu erstreckt, zeigte das höchste Ausmaß an Homologie mit dem Vertreter des lokalen Subtyps (B') (r142) (Figur 4 A). Darüber hinaus zeigte ein etwa 300 bp langer Sequenzbereich, der mit der 5'-Hälfte des Gens nef überlappt, höchste Homologie mit dem Subtyp (B')-Thai, wohingegen der verbleibende Teil einschließlich eines Fragments von 300 bp

Länge, der sich in die 3'-LTR-Region erstreckt, mit Subtyp C gruppiert (clustert) (Figur 4 B).

Unter Erweiterung der RIP-Analyse zeigten phylogenetischen Stammbäume die engste Verwandtschaft von vpr/vpu und dem 5'-Bereich des nef-Gens zu Isolaten des Subtyps B (Figur 5 A, B), wohingegen das 3'-nef-Fragment sich eindeutig mit Vertretern des Subtyps C gruppierte (Figur 5 C). Weitere Analysen bestätigten, daß die Sequenz des Subtyps B innerhalb dieses Mosaiks näher verwandt ist mit einem kürzlich beschriebenen Thai-(B')-Stamms (r142), der isoliert wurde von einem chinesischen IDU (Graf et al. 1998), als zu Prototyp-Isolaten des Subtyps B (mn und sf2) (Tabelle 2).

10

**Beispiel 13:**

Repräsentativer Charakter von 97cn54

In den kodierenden Regionen von vpr/vpu und dem nef-Gen von 97cn54 liegende Bruchstellen wurden in fast identischen Positionen bei allen Stämmen des Subtyps C, die aus in den nordwestlichen Provinzen von China lebenden IDUs isoliert wurden, gefunden. 2 RIP-Analysen, die repräsentativ für 8 unabhängig voneinander isolierte und analysierte HIV-1-Stämme von verschiedenen mit HIV-1 infizierten Personen in der autonomen Region Xinjiang isoliert wurden, sind in den Figuren 4 C und D dargestellt. Was die Herkunft von 97cn54 (Südwesten von China) und xj24 und xj15 (nordwestliches Gebiet) betrifft, legen diese Daten für die durch China zirkulierenden C/B'-rekombinanten Stämme einen gemeinsamen Vorläufer nahe. Unsere Ergebnisse zeigen also, daß 97cn54 ein C/(B')-Intersubtyp-Mosaikvirus mit 10 Bruchstellen der Intersubtyp-Rekombination darstellt, das unter den IDUs innerhalb der nordwestlichen Provinzen Chinas am stärksten prävalent ist. Eine schematische Darstellung des (B'/C)-Mosaikgenoms von Isolat 97cn545 ist in Figur 6 dargestellt.

25

**Beispiel 14:**

Vorhersage der über Subtypen hinaus kreuzreaktiven spezifischen Epitope für HIV-spezifische zytolytische T-Zellen

Genomische Sequenzen eröffnen die Möglichkeit, die Konserviertheit von bekannten CTL-Epitopen zu ermitteln, die einen Einfluß haben können auf die Effektivität von HIV-1-Impfstoffkandidaten. Die meisten Reagenzien und Daten bezüglich CTL-Epitopen stammen von Sequenzen von HIV-1<sub>LAI</sub> des Subtyps B. Um die Konserviertheit von über Subtypen hinaus

kreuzreaktiven CTL-Epitopen abzuschätzen, wurden die vorhergesagten Protein-Sequenzen von 97cn54 mit den bekannten und am besten kartierten LAI-spezifischen CTL-Epitopen verglichen. Von den 194 beschriebenen CTL-Epitopen von HIV-1 liegen 75, 55, 40 und 24 in Gag, (p17, p24, p15), in der Reversen Transkriptase (RT), in gp120 bzw. in gp41. Während fast 50% oder 5 mehr der Epitope in Gag und RT völlig identisch sind, stimmten nur 5% und 17% der von HIV-1<sub>LAI</sub> abgeleiteten CTL-Epitope von gp120 und gp41 exakt mit der für 97cn54 vorhergesagten Aminosäure-Sequenz überein. Wenn man jedoch zwei konservative Fehlpaarungen in einem bestimmten CTL-Epitop zuläßt, war ein zusätzlicher Bereich von 48% (p17), 33% (p24), 40% (RT), 57% (gp120) und 33% (gp41) der bekannten CTL-Epitope von HIV-1<sub>LAI</sub> verwandt mit den 10 Sequenzen in den entsprechenden von 97cn54 abgeleiteten Polypeptiden. Natürlich muß diese letzte Betrachtung mit einiger Vorsicht aufgenommen werden, da sogar nicht-konservative Austausche die HLA-Bindung oder die T-Zell-Rezeptorerkennung eines antigenen Peptids beseitigen kann. Zusammengenommen sagen diese Beobachtungen jedoch eindeutig eine beträchtliche über die Subtypen hinaus kreuzreaktive CTL-Reaktivität voraus, insbesondere der 15 funktionell und immunologisch konservierten Proteine von HIV-1. Außerdem legen diese Daten nahe, daß ein beträchtlicher Anteil der Reagenzien (Peptide, Vakziniavirus-Konstrukte), die für die Kartierung und Charakterisierung von CTL-Epitopen des Subtyps B synthetisiert und etabliert worden sind, auch nützlich sein können für die Bestimmung von CTL-Reaktivitäten auf Basis von HIV-Sequenzen des Subtyps C.

20

**Tabelle 3:** Leserahmen der kodierenden Sequenz von 97cn54

Leserahmen	Start	Ende	Start	Ende
gag	177	1654		
pol	1447	4458		
env	5589	8168		
vif	4403	4984		
vpr	4924	5214		
vpu	5426	5671		
tat	5195	5409	7730	7821
rev	5334	5409	7730	7821
nef	8170	8790		

Die Nummern beziehen sich auf das 5'-Ende der in SEQ ID NO: 1 wiedergegebenen DNA Sequenz.

**Beispiel 15:**

- 5 (A) Beschreibung der synthetischen kodierenden Region für C54gp160: C-gp160

Das C-gp120-Gen wurde in die einzigen KpnI/SacI-Restriktionsschnittstellen des pCR-Script amp(+) -Klonierungsvektors (Stratagene, Genbank Accession: U46017) kloniert. Die synthetische, im Kodongebrauch an stark exprimierte Säugergene angepaßte kodierende Region von C54gp160 ist in SEQ ID NO: 3 dargestellt. Die synthetische Signalsequenz kodiert ein

- 10 Transportsignal für den Import des kodierten Polypeptids in das endoplasmatische Retikulum.

Die Positionen der verschiedenen kodierenden Regionen sind wie folgt:

CDS	Start	Ende
synthetische Signalsequenz	28	87
gp160	88	2580

(B) Beschreibung der synthetischen Sequenz von C54 gagpolnef: C-gpnef

- 15 Das Gen C-gpnef wurde in die einzigen KpnI/SacI-Restriktionsschnittstellen des pCR-Script amp(+) Klonierungsvektors (Stratagene) kloniert. Die synthetische, im Kodongebrauch an stark exprimierte Säugergene angepaßte Sequenz von C54gagpolnef ist in SEQ ID NO: 2 dargestellt. In dem vorliegenden Konstrukt wurde das N-terminale Glycin gegen Alanin (Nukleotidsequenz GGC) ausgetauscht, um ein Targeting des Polypeptides an die Zytosplasmamembran und die
- 20 anschließende Sekretion von assemblierten Virus-ähnlichen Partikeln via Budding zu vermeiden. Gleichzeitig wurde an der natürlichen Frameshift-Sequenz ein (-1) Leserastersprung eingeführt, der ein obligates Durchlesen der Ribosomen aus dem Gag- in den Pol Leserahmen garantiert und so die Synthese eines GagPolNef Polyproteins sicherstellt.

Die Positionen der verschiedenen kodierenden Regionen sind wie folgt:

CDS	Start	Ende
gag	13	1500
5'pol ( $\Delta$ RT)	1501	2460
scrambled nef	2461	3090
3'pol ( $\Delta$ IN)	3091	4155
RT aktives Zentrum	4156	4266

Beispiel 16:

- 5 Das durch SEQ ID NO: 1 kodierte GagPolNef Polgen wurde über KpnI/XhoI in den Vektor pcDNA3.1 inseriert und in den E.coli Stamm XL1blue transformiert. Die Fähigkeit des GagPolNef Expressionsvektors eine Gag-spezifische Antikörperantwort zu induzieren wurde in weiblichen BALB/c Mäusen analysiert (Fig. 9). Zwei Gruppen von jeweils 5 Tieren erhielten jeweils eine intramuskuläre (i.m.) Primärimmunisierung von 100 µg DNA pro Immunisierung
- 10 gefolgt 2 i.m. Folgeimmunisierungen 3 und 6 Wochen später (Gruppe 1: pcDNA-GagPolNef; Gruppe 2: pcDNA). Eine Kontrollgruppe (Gruppe 3) wurde lediglich mit PBS immunisiert. Die Gesamttiter an Gag-spezifischem IgG wurden gegen gereinigtes Gag-Protein im ELISA bestimmt. Die Impfung mit pcDNA-GagPolNef resultierte in einer schnellen Induktion hoher Titer an Gag spezifischen Antikörpern (1:4.000), die gekennzeichnet war durch ein typisches
- 15 Th1 Profil an Antikörper Isotypen (IgG2a >> IgG1). Die beiden Kontrollgruppen 2 und 3 lieferten keine Hinweise auf die Generierung Gag-spezifischer Antikörper. Die Antikörpertiter stiegen beinahe um das hundertfache (1:20.000) 1 Woche nach der ersten Folgeimmunisierung und erreichten Gag-spezifische Endpunkttiler von 1:80,000 eine Woche nach der zweiten Boosterimmunisierung. Zu keinem Zeitpunkt konnte bei den beiden Kontrollgruppen eine
- 20 signifikante, Gag-spezifische Antikörperantwort nachgewiesen werden.

**Beispiel 17:**

- Die Antigen-spezifische Zytokinsekretion als Hinweis auf die Induktion einer T-Helfer *Memory*-Antwort wurde aus Milzzellen analysiert, die jeweils 5 Tage nach der zweiten Folgeimmunisierung entnommen wurden. Die Milzzellen der Mäuse, die drei i.m.
- 5    Immunisierungen mit pcDNA-GagPolNef erhalten hatten, reagierten mit einer deutlichen gIFN Sekretion auf Gag-spezifischen Antigenstimulus (Tabelle 3). Eine vergleichsweise reduzierte gIFN Produktion wurde Milzzellen beobachtet, die aus Mäusen nach dreimaliger subkutaner (s.c.) oder intradermaler (i.d.) Immunisierung mit pcDNA-GagPolNef nach dem selben Schema wie oben gewonnen wurden. In allen Immunisierungsgruppen wurden, unabhängig von der
- 10    Immunisierungsroute, keine nennenswerte IL4- und IL5 Sekreten aus den spezifisch *in vitro* restimulierten Milzzellen festgestellt. Eine Zytokinsekretion aus nicht-stimulierten Milzzellen wurde nicht beobachtet.

Die i.m. Immunisierung mit pcDNA-GagPolNef führte demnach zu einem starken Th1 Zytokin-Profil, während die subkutane Verabreichung eher eine schwache Th1 Antwort induzierte.

**Tabelle 4:** Zytokin-Profil von *in vitro* Gag-stimulierten Milzzellen von Mäusen i Immunisierung (Nadelinjektion) oder *i.d.* bzw. *s.c.* Immunisierung durch eine Part mit den angegebenen DNA Konstrukten

DNA Vakzine	IL-4 (pg/ml)	IL-5 (pg/ml)	IFN- $\gamma$ (pg/ml)
pcDNA-GagPolNef (i.m.)	<8	<16	3220 $\pm$ 840
pcDNA-GagPolNef (i.d.)	<8	<16	80 $\pm$ 32
pcDNA-GagPolNef (s.c.)	<8	<16	<32

Mittelwerte  $\pm$  Standardabweichung von Milzzellen, gewonnen jeweils aus 5 Mäusen pro Experiment

**Beispiel 18:**

- 20    Um die Fähigkeit von pcDNA-GagPolNef zur Induktion Gag-spezifischer CTLs zu überprüfen wurden Milzzellen 3 Wochen nach einer primären Immunisierung mit pcDNA-GagPolNef (Gruppe 1), pcDNA (Gruppe 2) und PBS (Gruppe 3) *in vitro* in einer gemischen-Lymphozyten-Tumor-Zellkultur für 6 Tage spezifisch restimuliert und anschließend hinsichtlich ihrer zytotoxischen Aktivität untersucht. Bei dem nonameren, vom Gag Protein der Subtyp B-Viren

(IIIb-Isolat) abgeleiteten AMQMLKETI Peptid (Einbuchstabencode), das in diesem Versuch zur *in vitro* Restimulation gleichwie zur Bestimmung der spezifischen zytotoxischen Aktivität eingesetzt wurde, stellt bekannterweise ein D<sup>d</sup>-restringiertes CTL Epitope in der BALB/c Maus dar. *Gag*-spezifische zytotoxische T-Zellen konnten nach einer einmaligen i.m. Injektion mit dem pcDNA-GagPolNef Plasmid, nicht jedoch in einer der beiden Kontrollgruppen 2 und 3 festgestellt werden. Die Behandlung von Milzzellen mit dem oben genannten Peptid resultierte nicht in einem *in vitro Priming* *Gag*-spezifischer zytotoxischer T-Zellen. Diese Ergebnisse bestätigen (i) die Fähigkeit von pcDNA-GagPolNef zur Induktion Spezifischer zytotoxischer T-Zellen, die (ii) Subtyp-übergreifend aktiv sind (Figur 9).

10

#### Literatur

- Bai, X., Su, L., Zhang, Y., and et al(1997). Subtype and sequence analysis of the C2V3 region of gp120 gene among HIV-1 strains in Xinjiang. *Chin. J. Virology* 13.
- Carr, J. K., Salminen, M. O., Koch, C., Gotte, D., Artenstein, A. W., Hegerich, P. A., St Louis, D., Burke, D. S., and McCutchan, F. E.(1996). Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand. *J. Virol.* 70, 5935-5943.
- Carr, J. K., Salminen, M. O., Albert, J., Sanders Buell, E., Gotte, D., Birx, D. L., and McCutchan, F. E.(1998). Full genome sequences of human immunodeficiency virus type 1 subtypes G and A/G intersubtype recombinants. *Virology* 247, 22-31.
- Esparza, J., Osmanov, S., and Heyward, W. L.(1995). HIV preventive vaccines. Progress to date. *Drugs* 50, 792-804.
- Expert group of joint United Nations programme on HIV/AIDS(1999). Implications of HIV variability for transmission: scientific and policy issues. *AIDS* 11, UNAIDS 1-UNAIDS 15.
- Gao, F., Robertson, D. L., Morrison, S. G., Hui, H., Craig, S., Decker, J., Fultz, P. N., Girard, M., Shaw, G. M., Hahn, B. H., and Sharp, P. M.(1996). The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. *J. Virol.* 70, 7013-7029.
- Gao, F., Robertson, D. L., Carruthers, C. D., Morrison, S. G., Jian, B., Chen, Y., Barre Sinoussi, F., Girard, M., Srinivasan, A., Abimiku, A. G., Shaw, G. M., Sharp, P. M., and Hahn, B. H.(1998). A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. *J. Virol.* 72, 5680-5698.
- Gaywee, J., Artenstein, A. W., VanCott, T. C., Trichavaroj, R., Sukchamnong, A., Amlee, P.,

- de Souza, M., McCutchan, F. E., Carr, J. K., Markowitz, L. E., Michael, R., and Nittayaphan, S.(1996). Correlation of genetic and serologic approaches to HIV-1 subtyping in Thailand. *J. Acquir. Immune. Defic. Syndr. Hum. Retrovirol.* 13, 392-396.
- Graf, M., Shao, Y., Zhao, Q., Seidl, T., Kostler, J., Wolf, H., and Wagner, R.(1998). Cloning and characterization of a virtually full-length HIV type 1 genome from a subtype B'-Thai strain representing the most prevalent B-clade isolate in China. *AIDS Res. Hum. Retroviruses* 14, 285-288.
- Graham, B. S. and Wright, P. F.(1995). Candidate AIDS vaccines. *N. Engl. J. Med.* 333, 1331-1339.
- 10 Kostrikis, L. G., Bagdades, E., Cao, Y., Zhang, L., Dimitriou, D., and Ho, D. D.(1995). Genetic analysis of human immunodeficiency virus type 1 strains from patients in Cyprus: identification of a new subtype designated subtype I. *J. Virol.* 69, 6122-6130.
- Leitner, T. and Albert, J.(1995). *Human Retroviruses and AIDS 1995: a compilation and analysis of nucleic acid and amino acid sequences.* (Myers, G., Korber, B., Wain-Hobson, S.,
- 15 Jeang, K., Mellors, J., McCutchan, F., Henderson, L., and Pavlakis, G. Eds.) Los Alamos National Laboratory, Los Alamos, N. Mex. III147-III150.
- Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., Ingersoll, R., Sheppard, H. W., and Ray, S. C.(1999). Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* 73, 152-160.
- 20 Loussert Ajaka, I., Chaix, M. L., Korber, B., Letourneur, F., Gomas, E., Allen, E., Ly, T. D., Brun Vezinet, F., Simon, F., and Saragosti, S.(1995). Variability of human immunodeficiency virus type 1 group O strains isolated from Cameroonian patients living in France. . *J. Virol.* 69, 5640-5649.
- 25 Luo, C. C., Tian, C., Hu, D. J., Kai, M., Dondero, T., and Zheng, X.(1995). HIV-1 subtype C in China [letter]. *Lancet* 345, 1051-1052.
- Myers, G., Korber, B., Foley, B., Jeang, K. T., Mellors, J. W., and Wain Hobson, S.(1996). *Human retroviruses and AIDS: a compilation and analysis of nucleic acid and amino acid sequences.* (Anonymous Theoretical Biology and Biophysics Group, Los Alamos, N. Mex.
- 30 Salminen, M. O., Koch, C., Sanders Buell, E., Ehrenberg, P. K., Michael, N. L., Carr, J. K., Burke, D. S., and McCutchan, F. E.(1995). Recovery of virtually full-length HIV-1 provirus of diverse subtypes from primary virus cultures using the polymerase chain reaction. *Virology* 213,

80-86.

Shao, Y., Zhao, Q., Wang B., and et al(1994). Sequence analysis of HIV env gene among HIV infected IDUs in Yunnan epidemic area of China. *Chin. J. Virology* **10**, 291-299.

Shao, Y., Su, L., Sun, X., and et al(1998). Molecular Epidemiology of HIV infection in China.

5 *12th world AIDS conference, Geneva* **13132**, (Abstract).

Shao, Y., Guan, Y., Zhao, Q., and et al(1999). Genetic variation and molecular epidemiology of the Ruily HIV-1 strains of Yunnan in 1995. *Chin. J. Virol.* **12**, 9.

Sharp, P. M., Robertson, D. L., and Hahn, B. H.(1995). Cross-species transmission and recombination of 'AIDS' viruses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **349**, 41-47.

10 Sharp, P. M., Bailes, E., Robertson, D. L., Gao, F., and Hahn, B. H.(1999). Origins and evolution of AIDS viruses. *Biol. Bull.* **196**, 338-342.

World Health Organisation Network for HIV Isolation and Characterization(1994). HIV-1 variation in WHO-sponsored vaccine-evaluation sites:genetic screening, sequence analysis and preliminary biological characterization of selected viral strains. *AIDS Res. Hum. Retroviruses*

15 **10**, 1327-1344.

Yu, H., Su, L., and Shao, Y.(1997). Identification of the HIV-1 subtypes by HMA and sequencing. *Chin. J. Epidemiol.* **18**, 201-204.

**Patentansprüche**

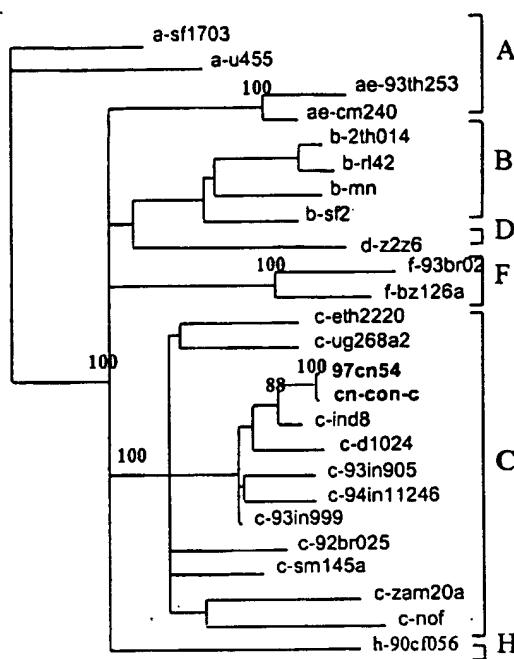
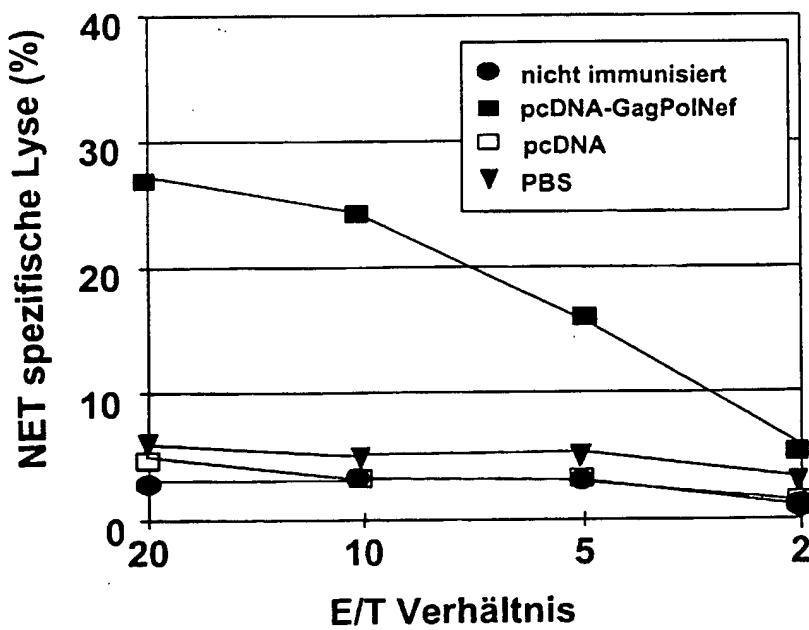
1. Ein Polynukleotid umfassend eine Nukleinsäuresequenz gemäß SEQ ID NO:1, 2 oder 3 oder dessen Fragment oder Derivat, oder ein Polynukleotid, das mit der Nukleinsäuresequenz gemäß SEQ ID NO:1, 2 oder 3 hybridisiert.  
5
2. Polynukleotid oder dessen Fragment oder Derivat nach Anspruch 1, wobei das hybridisierende Polynukleotid unter stringenten Bedingungen mit der Nukleinsäuresequenz gemäß SEQ ID NO:1, 2 oder 3 hybridisiert.  
10
3. Polynukleotid oder dessen Fragment oder Derivat nach Anspruch 1 oder 2, umfassend mindestens eine kontinuierliche Sequenz von mindestens 9 Nukleotiden, bevorzugt mindestens 15, mehr bevorzugt mindestens 27, oder mehr Nukleotide.  
15
4. Polynukleotid oder dessen Fragment oder Derivat nach Anspruch 3 mit mehr als einer kontinuierlichen Sequenz von Nukleotiden, wobei mindestens zwei der kontinuierlichen Sequenzen durch einen Nukleotid-Platzhalter ("spacer") getrennt sind.  
20
5. Polynukleotid oder dessen Fragment oder Derivat nach einem der Ansprüche 1 bis 4, welches für wenigstens ein Polypeptid kodiert, welches kodiert wird durch die unter SEQ ID NO: 1, SEQ ID NO: 2 oder SEQ ID NO: 3 beschriebene Nukleotidsequenz.  
25
6. DNA-Konstrukte, umfassend das Polynukleotid oder dessen Fragment oder Derivat nach einem der Ansprüche 1 bis 5.  
30
7. Bakterieller oder viral Vektor, umfassend das Polynukleotid oder dessen Fragment oder Derivat nach einem der Ansprüche 1 bis 5.
8. Polynukleotid oder dessen Fragment oder Derivat nach einem der Ansprüche 1 bis 5 als Arzneimittel, Impfstoff oder Diagnostikum.  
35
9. Verwendung des Polynukleotids oder dessen Fragments oder Derivats nach einem der

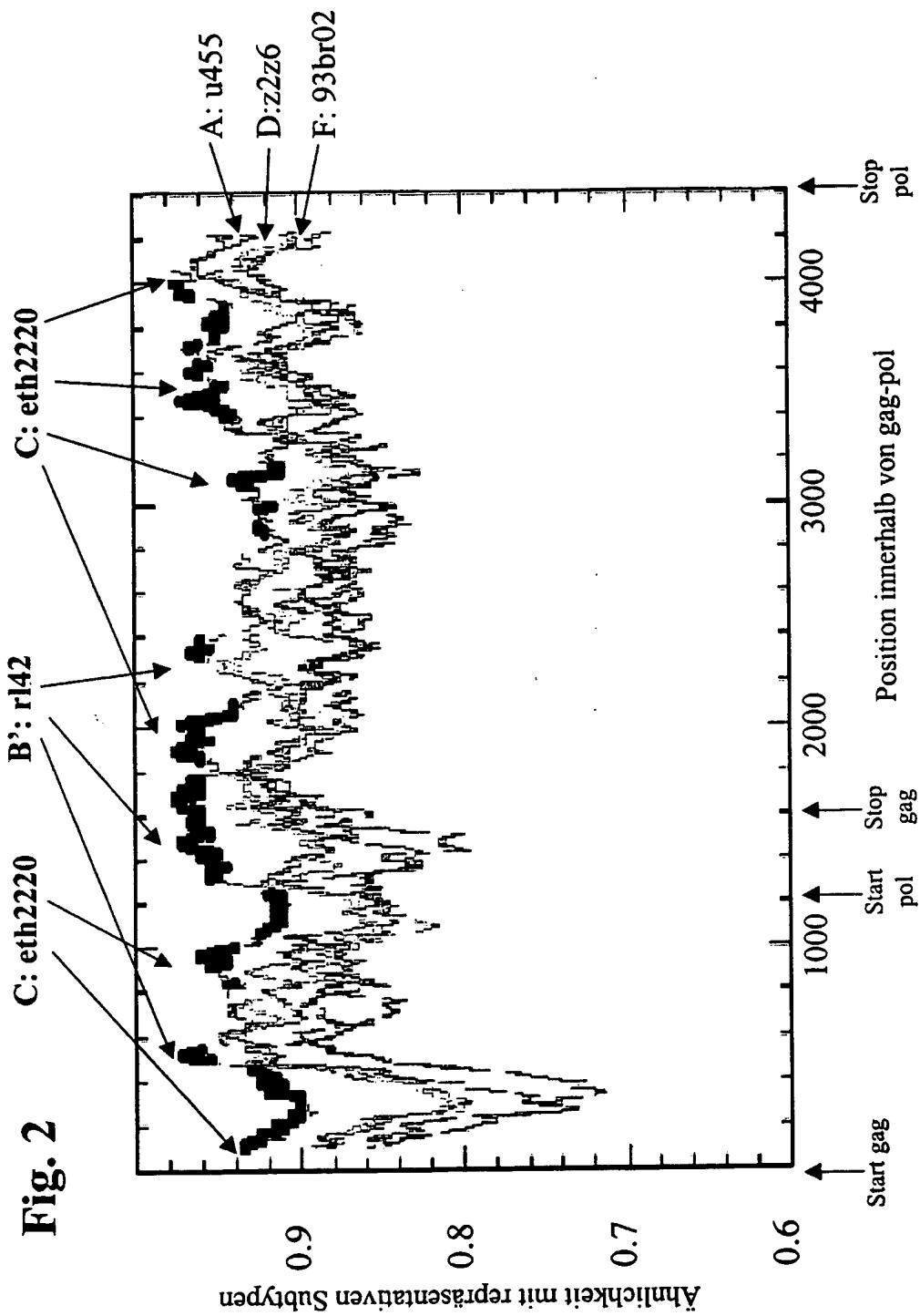
Ansprüche 1 bis 5 zur Herstellung eines Arzneimittels oder Impfstoffs für die Behandlung oder Prävention von HIV-Infektionen.

10. 10. Polypeptid, kodiert von der Nukleotidsequenz oder Fragment oder Derivat der  
5 Nukleotidsequenz gemäß SEQ ID NO:1, 2 oder 3.
11. 11. Polypeptid nach Anspruch 10, umfassend eine kontinuierliche Sequenz von von  
mindestens 8 Aminosäuren, die von der Nukleotidsequenz oder Fragmenten oder  
Derivaten der Nukleotidsequenz gemäß SEQ ID NO:1, 2 oder 3 kodiert werden.  
10
12. 12. Polypeptid nach Anspruch 10 oder 11, wobei die Aminosäuresequenz dem HIV-  
Hüllprotein oder einem Fragment des HIV-Hüllproteins entspricht.  
15
13. 13. Polypeptid nach einem der Ansprüche 10 bis 12, ferner umfassend eine antigene  
Determinante, die natürlicherweise in Infizierten eine Immunreaktion auslöst.  
15
14. 14. Polypeptid nach Anspruch 13, wobei die antigene Determinante ein Konformations-  
Epitop oder ein lineares Epitop ist.  
20
15. 15. Das Polypeptid nach einem der Ansprüche 10 bis 14 als Arzneimittel, Impfstoff oder  
Diagnostikum.  
25
16. 16. Verwendung des des Polypeptids nach einem der Ansprüche 10 bis 14 zur Herstellung  
eines Arzneimittels oder Impfstoffs für die Behandlung oder Prävention von HIV-  
Infektionen.  
25
17. 17. Isoliertes Polypeptid spezifisch gegen ein Polypeptid nach einem der Ansprüche 10 bis  
14.  
30
18. 18. Isoliertes Polypeptid nach Anspruch 17 als Arzneimittel oder Diagnostikum.  
30
19. 19. Verwendung des isolierten Polypeptids nach Anspruch 17 zur Herstellung eines

Arzneimittels für die Behandlung oder Prävention von HIV-Infektionen.

20. Isoliertes Polypeptid nach Anspruch 17 oder 18 oder die Verwendung des isolierten Polypeptids nach Anspruch 19, wobei das isolierte Polypeptid ein Antikörper ist.

**Fig. 1****Fig. 9**



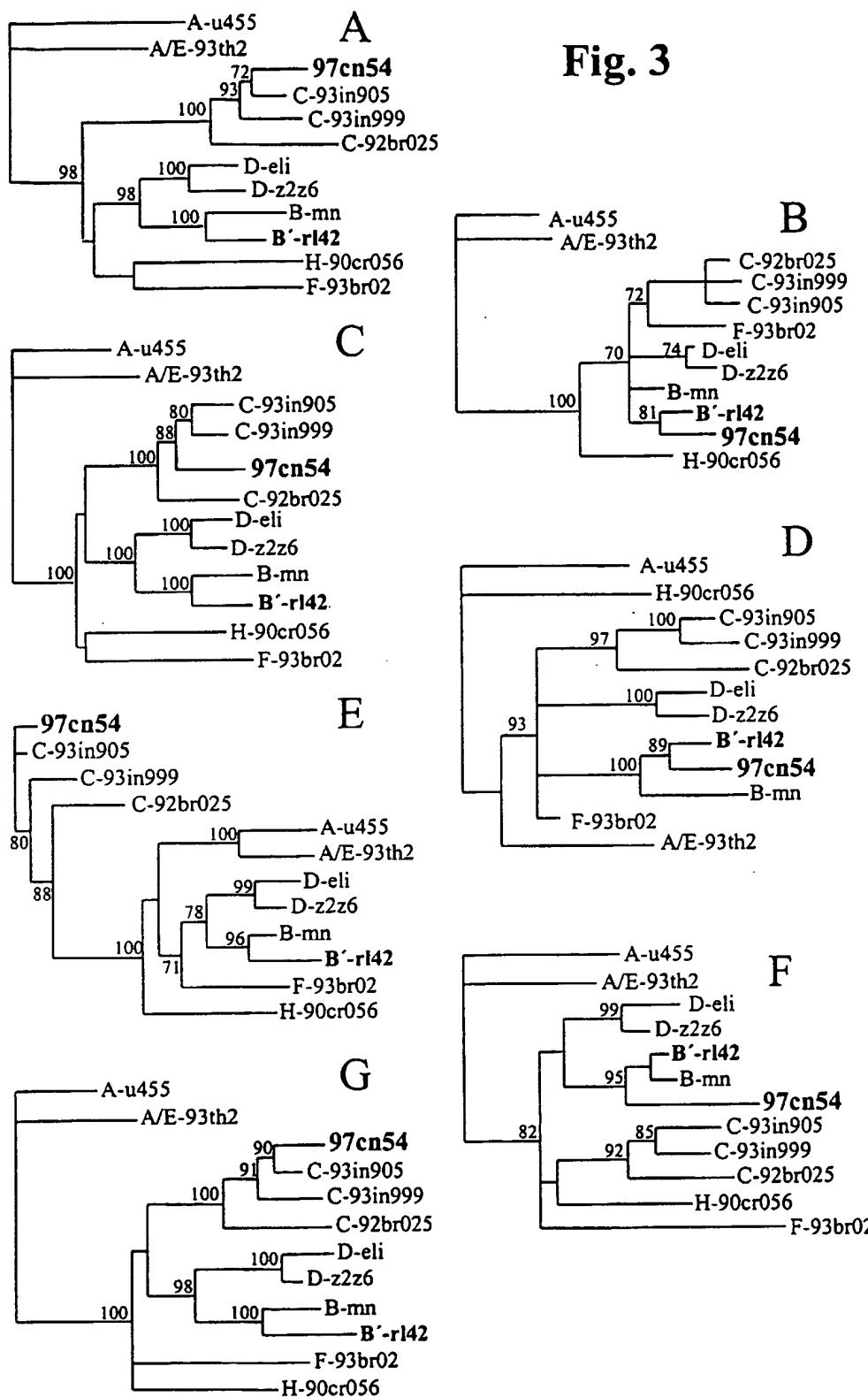
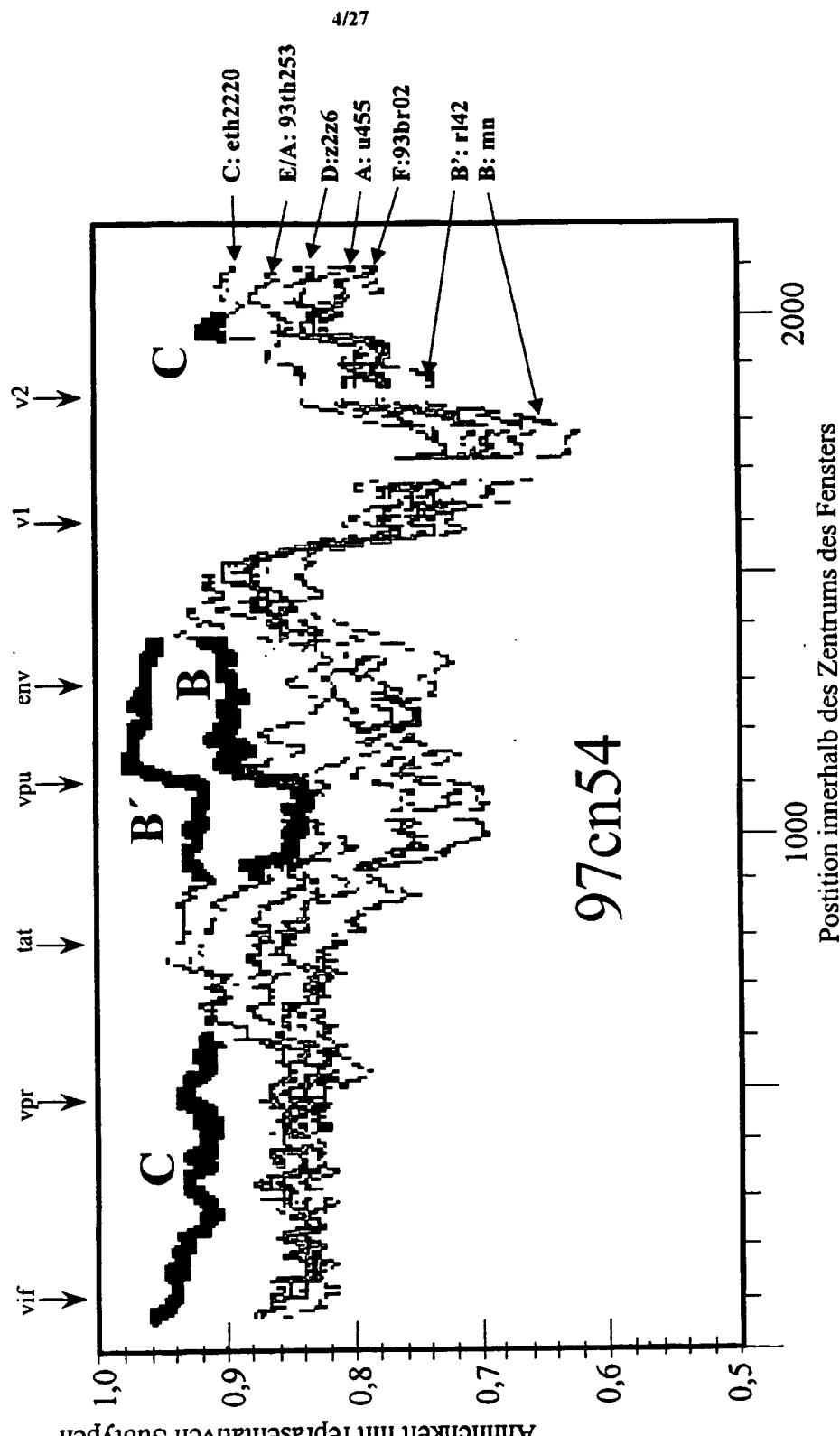
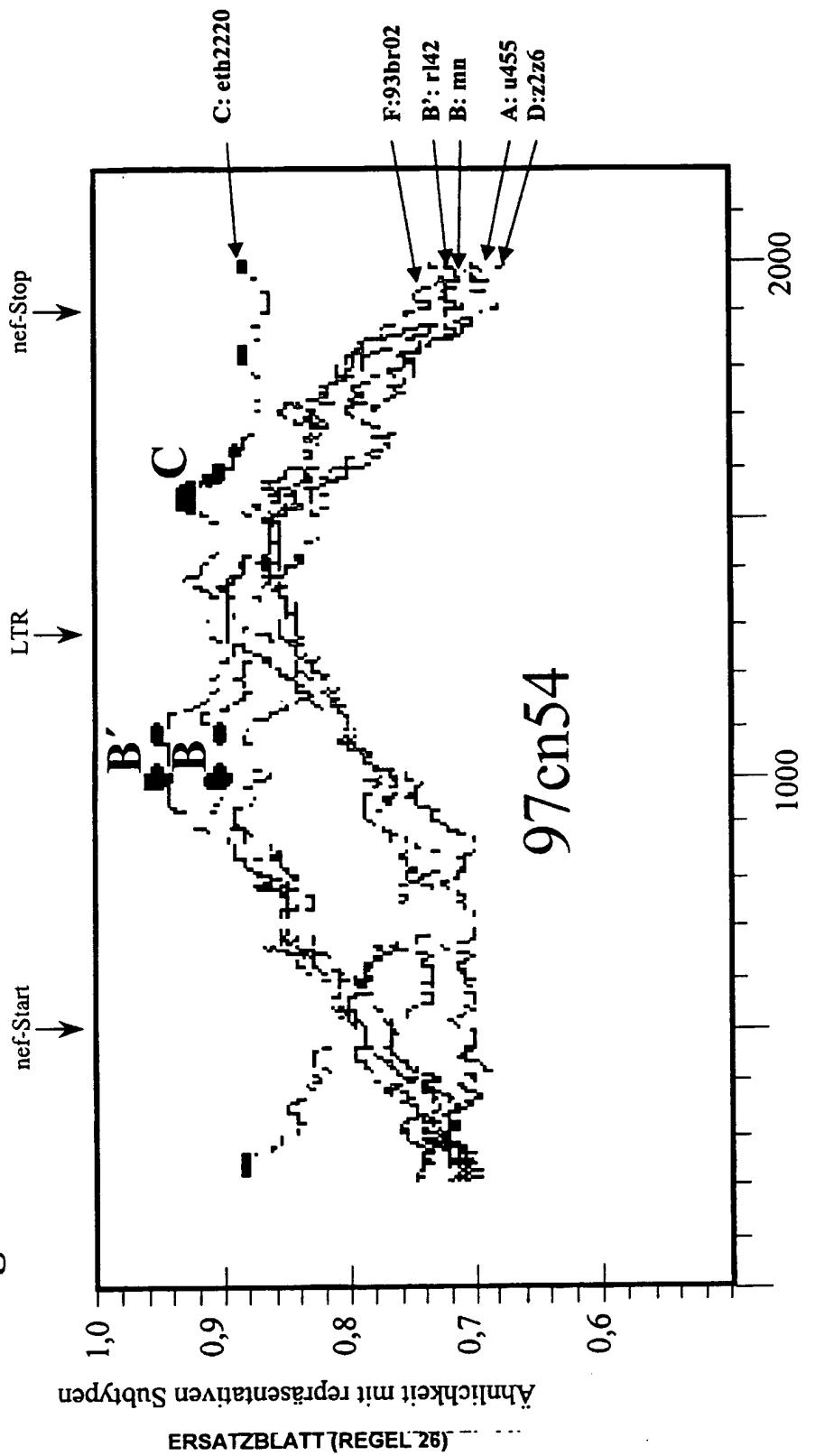
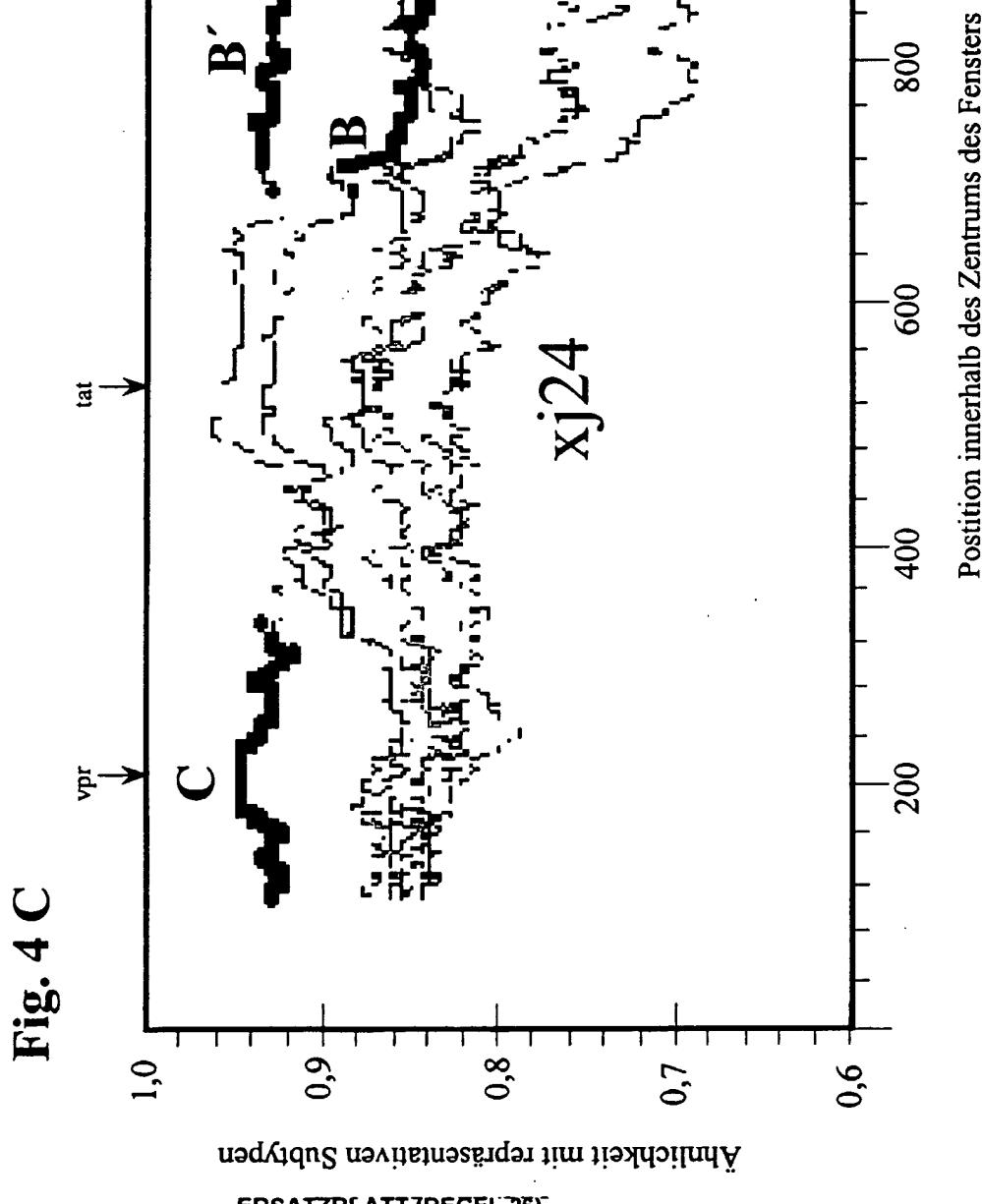
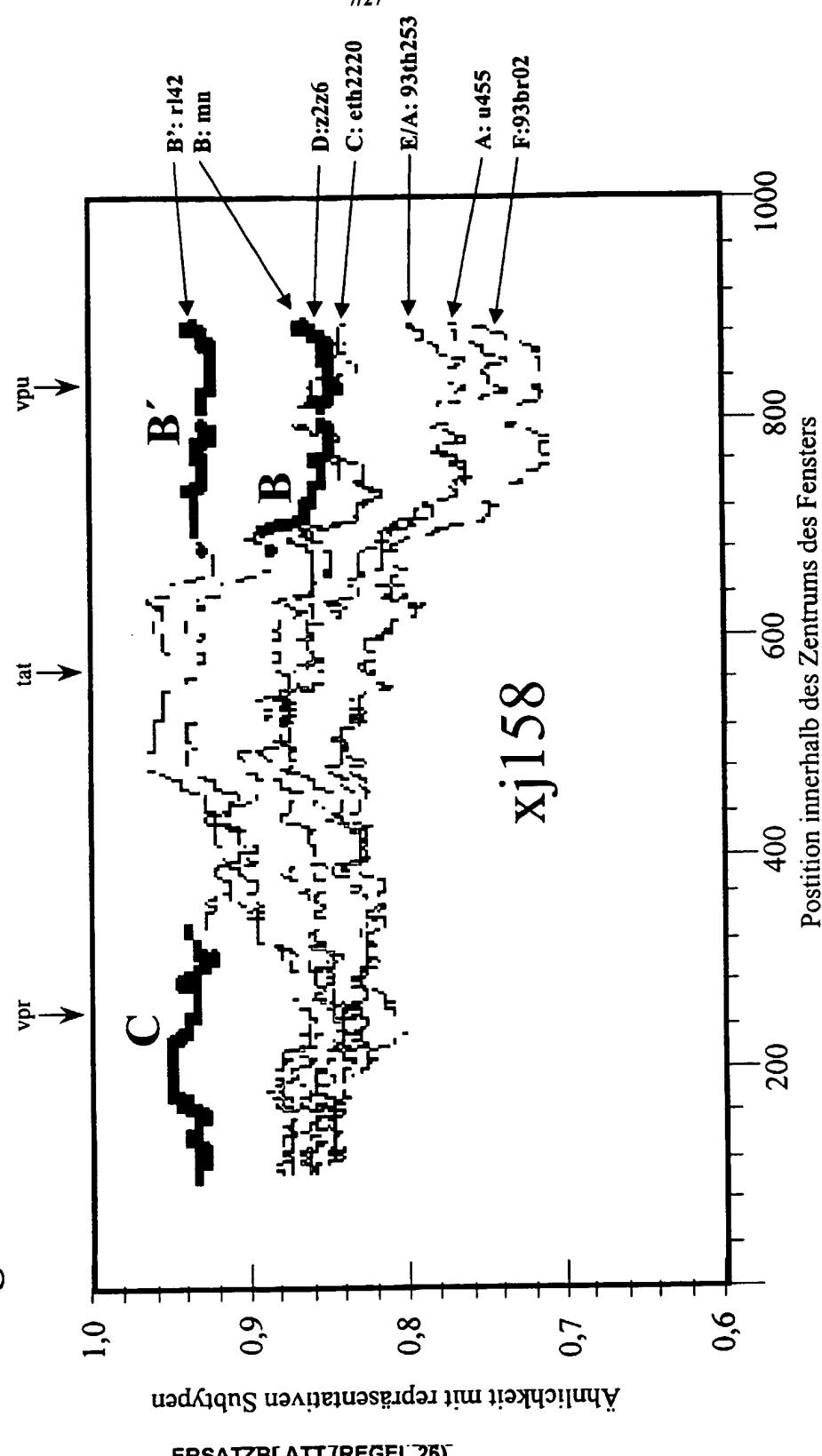


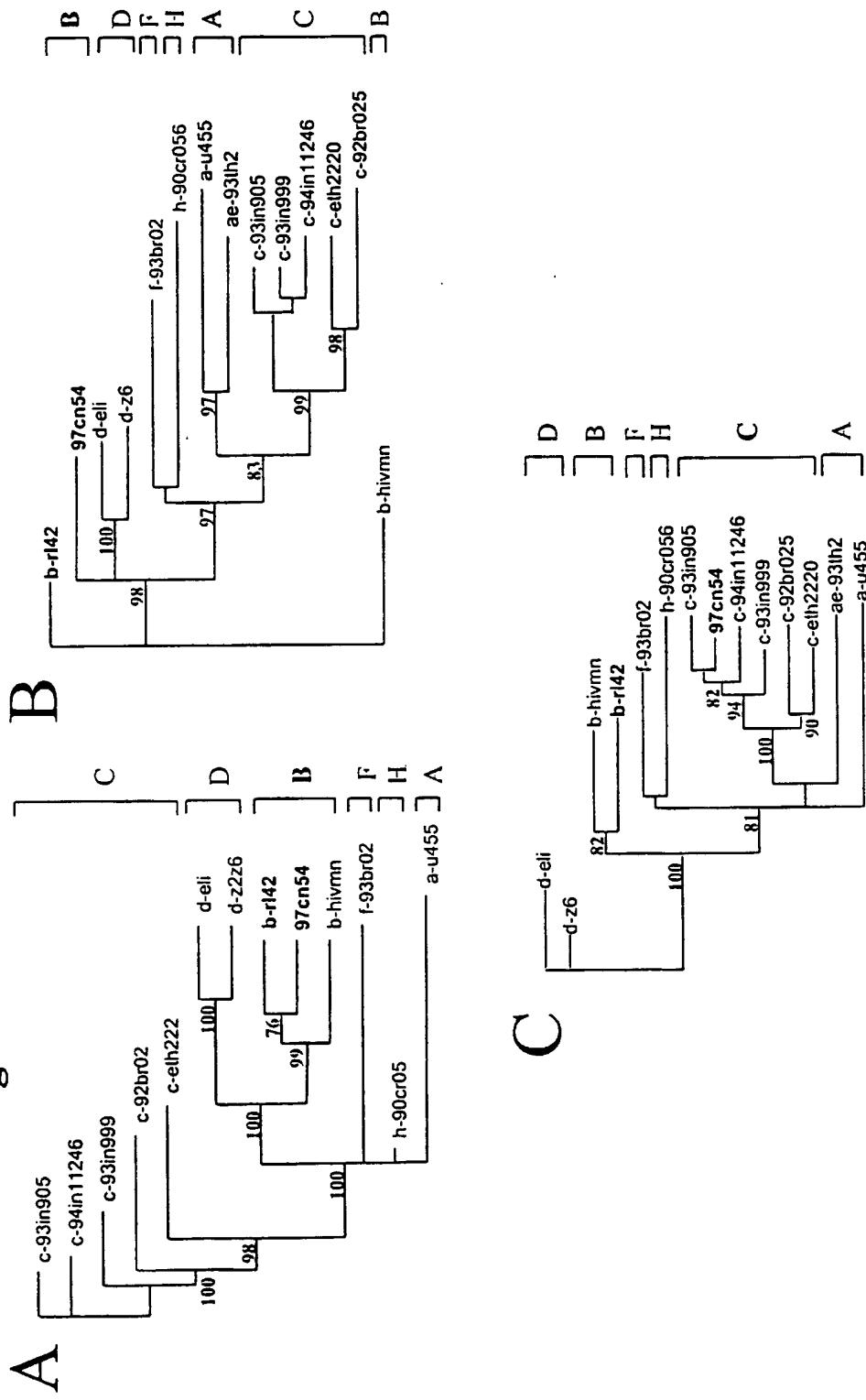
Fig. 3

**Fig. 4 A**

**Fig. 4 B**

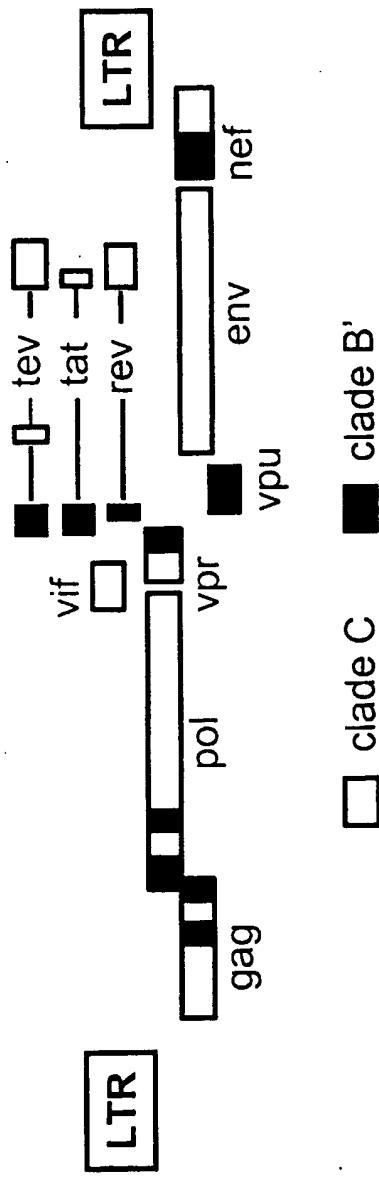


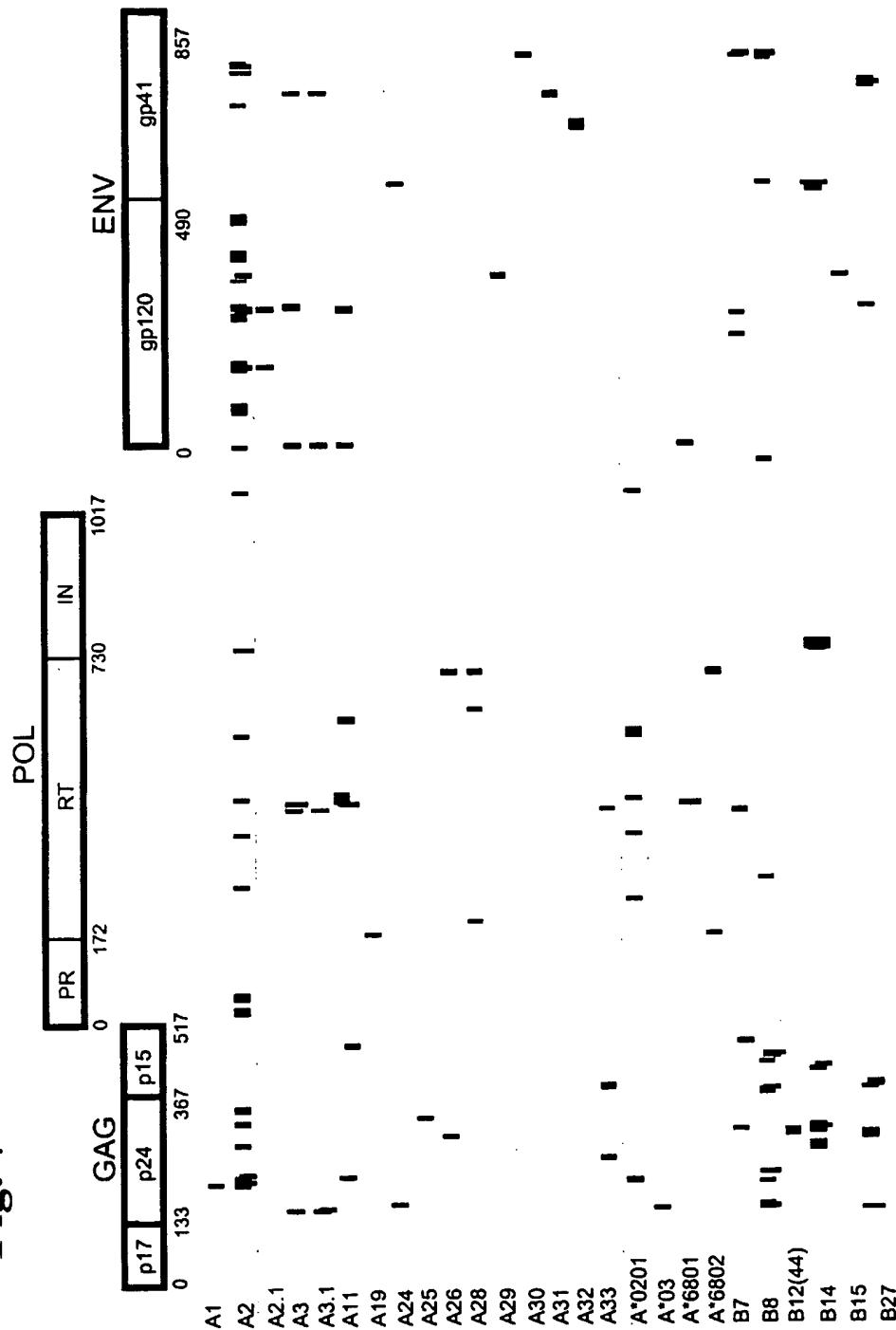
**Fig. 4 D**

**Fig. 5**

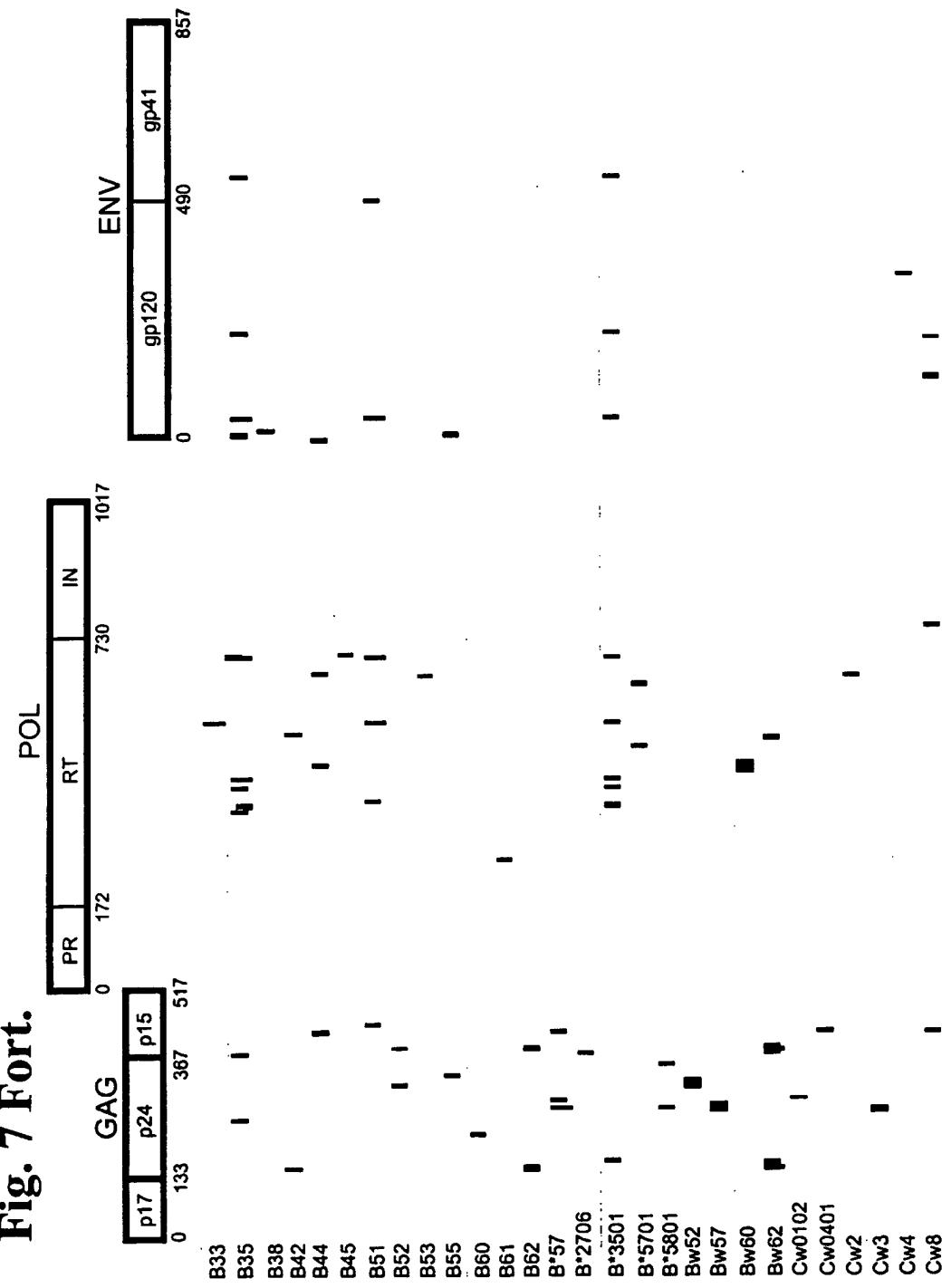
**Fig. 6**

## Mosaikartige Organisation des Genoms von 97cn54



**Fig. 7**

ERSATZBLATT (REGEL 26)

**Fig. 7 Fort.**

**Fig. 8/a**

AATCTCTAGCAGTGGCGCCGAACAGGGACTTGAAAGCGAAAGTAAGACCAGAGGAGATC  
 1 - +-----+-----+-----+-----+-----+-----+-----+-----+ 60  
 TTAGAGATCGTCACCGCGGGCTGTCCCTGAACCTTCGCTTCATTCTGGCTCCTCTAG  
 a N L \* Q W R P N R D L K A K V R P E E I -  
 b I S S S G A R T G T \* K R K \* D Q R R S -  
 c S L A V A P E Q G L E S E S K T R G D L -  
  
 TCTCGACGCAGGACTCGGCTTGCTGAAGTCACCGCAAGAGGCCAGAGCGGGCGACTGG  
 61 -----+-----+-----+-----+-----+-----+-----+-----+ 120  
 AGAGCTCGCTCTGAGGCCAACGACTTCACGTGAGCCGTTCCGCTCGCCGCTGACC  
 a S R R R T R L A E V H S A R G E S G D W -  
 b L D A G L G L L K C T R Q E A R A A T G -  
 c S T Q D S A C \* S A L G K R R E R R L V -  
  
 TGAGTACGCCAATTATATTGACTAGCGGAGGCTAGAAGGAGAGAGATGGGTGCGAGAGC  
 121 -----+-----+-----+-----+-----+-----+-----+-----+ 180  
 ACTCATCGGGTTAATATAAACTGATCGCCCTCGATCTTCCCTCTCTACCCACGCTCTCG  
 a \* V R Q L Y L T S G G \* K E R D G C E S -  
 b E Y A N Y I \* L A E A R R R E M G A R A -  
 c S T P I I F D \* R R L E G E R W V R E R -  
  
 GTCAATATTAAGGGGGAAAATTAGATAATGGGAAAAATTAGGTTAAGGCCAGGGGG  
 181 -----+-----+-----+-----+-----+-----+-----+-----+ 240  
 CAGTTATAATTCTCCCCCTTTAATCTATTACCCCTTTAATCCAATTCCGTCCCCC  
 a V N I K R G K I R \* M G K N \* V K A R G -  
 b S I L R G G K L D K W E K I R L R P G G -  
 c Q Y \* E G E N \* I N G K K L G \* G Q G E -  
  
 AAAGAAAACACTATATGCTAAAACACCTAGTATGGGCAAGCAGGGAGCTGGAAAGATTG  
 241 -----+-----+-----+-----+-----+-----+-----+-----+ 300  
 TTTCTTGTGATATACGATTGTGGATCATACCCGTTCTGCTCCCTCGACCTTCTAAACG  
 a K E T L Y A K T P S M G K Q G A G K I C -  
 b K K H Y M L K H L V W A S R E L E R F A -  
 c R N T I C \* N T \* Y G Q A G S W K D L H -  
  
 ACTTAACCTGGCTTTAGAGACATCAGAAGGCTGAAACAAATAATGAAACAGCTACA  
 301 -----+-----+-----+-----+-----+-----+-----+-----+ 360  
 TGAATTGGACCGGAAAATCTGTAGTCTTCGACATTGTTATTACTTTGTCGATGT  
 a T \* P W P F R D I R R L \* T N N E T A T -  
 b L N P G L L E T S E G C K Q I M K Q L Q -  
 c L T L A F \* R H Q K A V N K \* \* N S Y N -  
  
 ATCAGCTTCAGACAGGAACAGAGGAACCTAGATCATTATTCAACACAGTAGCAACTCC  
 361 -----+-----+-----+-----+-----+-----+-----+-----+ 420  
 TAGTCGAGAAGTCTGCTCTTGCTCTTGAAATCTAGTAATAAGTTGTCATCGTTGAGG  
 a I S S S D R N R G T \* I I I Q H S S N S -  
 b S A L Q T G T E E L R S L F N T V A T P -  
 c Q L F R Q E Q R N L D H Y S T Q \* Q L P -  
  
 CTATTGTGTACATACAGAGATAGATGTACGAGACACCGAGAGAAGCCTTAGACAAGATAGA  
 421 -----+-----+-----+-----+-----+-----+-----+-----+ 480  
 GATAACACATGTATGCTCTATCTACATGCTCTGTGGCTCTCGGAATCTGTTCTATCT  
 a L L C T Y R D R C T T R H Q R S L R Q D R -  
 b Y C V H T E I D V R D T R E A L D K I E -  
 c I V Y I Q R \* M Y E T P E K P \* T R \* R -  
  
 GGAAGAACAAAACAAAATTAGCAAAACACAGCAGGCCAAGGGAGGCTGACGGGAAGGT  
 481 -----+-----+-----+-----+-----+-----+-----+-----+ 540  
 CCTTCTTGTGTTGTTAAGTCGTTTGTGTCGTCGTTCTCCGACTGCCCTCCA  
 a G R T K Q N S A K N T A G K G G \* R E G -  
 b E E Q N K I Q Q K T Q Q A K E A D G K V -  
 c K N K T K F S K K H S R Q R R L T G R S -

**Fig. 8/b**

CAGTCAAAATTATCCTATAAGTACAGAATCTCAAGGGCAAATGGTACATCAGCCCATATC  
 541 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+ 600  
 GTCAGTTTAATAGGATATCATGTCTTAGGTTCCCGTTACCATGTAGTCGGTATAG  
 a Q S K L S Y S T E S P R A N G T S A H I -  
 b S Q N Y P I V Q N L Q G Q M V H Q P I S -  
 c V K I I L \* Y R I S K G K W Y I S P Y H -  
  
 ACCTAGAACTTTAACATGGTAAAGTGGTAGAAGAGAAGGCTTTAGCCCAGAAGT  
 601 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+ 660  
 TGGATCTTGAATTTACGTACCCATTTCACCATCTTCTCCGAAATCGGGCTTCA  
 a T \* N F K C M G K S G R R E G F \* P R S -  
 b P R T L N A W V K V V E E K A F S P E V -  
 c L E L \* M H G \* K W \* K R R L L A Q K \* -  
  
 AATAACCCATGTTTCAGCGTTATCAGAAGGAGCCACCCACAAGATTAAACACCATGCT  
 661 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+ 720  
 TTATGGGTACAAAGTCGCAATAGTCTTCCCGTGGGTGTTCTAAATTGTTGACGA  
 a N T H V F S V I R R S H P T R F K H H A -  
 b I P M F S A L S E G A T P Q D L N T M L -  
 c Y P C F Q R Y Q K E P P H K I \* T P C \* -  
  
 AAACACAGTGGGGGACATCAAGCAGCTATGCAAATATTAAAGATAACCATCAATGAAGA  
 721 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+ 780  
 TTTGTGTCACCCCCTGTAGTCGTCGATACGTTATAATTCTATGGTAGTTACTTCT  
 a K H S G G T S S S Y A N I K R Y H Q \* R -  
 b N T V G G H Q A A M Q I L K D T I N E E -  
 c T Q W G D I K Q L C K Y \* K I P S M K R -  
  
 GGCTGCAGAATGGGATAGATTACATCCAGTACATGCAGGGCTATTGCACCAGGCCAAT  
 781 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+ 840  
 CCGACGTCTTACCCCTATCTAATGTAGGTCTGACGTACGTCGGATAACGTGGTCCGGTTA  
 a G C R M G \* I T S S T C R A Y C T R P N -  
 b A A E W D R L H P V H A G P I A P G Q M -  
 c L Q N G I D Y I Q Y M Q G L L H Q A K \* -  
  
 GAGAGAACCAAGGGAAAGTGACATAGCAGGAACACTACTAGAACCTACAGAACAAATAGC  
 841 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+ 900  
 CTCTCTTGGTCCCTTCACTGTATCGTCCTGATGATCATTGGATGTCCTGTTATCG  
 a E R T K G K \* H S R N Y \* \* P T G T N S -  
 b R E P R G S D I A G T T S N L Q E Q I A -  
 c E N Q G E V T \* Q E L L V T Y R N K \* H -  
  
 ATGGATGACGGAGTAACCCACCTGTTCCAGTAGGAGACATCTATAAAAGATGGATAATTCT  
 901 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+ 960  
 TACCTACTGCTCATTGGGGACAAGGTCTGATCCTCTGTAGATATTCTACCTATTAAGA  
 a M D D E \* P T C S S R R H L \* K M D N S -  
 b W M T S N P P V P V G D I Y K R W I I L -  
 c G \* R V T H L F Q \* E T S I K D G \* F W -  
  
 GGGATTAAATAAAATAGTAAGAATGTATAGCCCTACCAGCATTCTGGACATAAAACAAGG  
 961 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+ 1020  
 CCCTAATTATTTATCATCTTACATATCGGGATGGTCGAAGACCTGTATTGTTCC  
 a G I K \* N S K N V \* P Y Q H S G H K T R -  
 b G L N K I Y R M Y S P T S I L D I K Q G -  
 c D \* I K \* \* E C I A L P A F W T \* N K G -  
  
 GCCAAAGGAACCTTGTAGAGACTATGTAGACCGGTTCTTAAACTTAAGAGCGAAC  
 1021 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+ 1080  
 CGGTTTCTGGAAATCTGTATACATCTGGCCAAGAAATTGAAATTCTCGCCTTGT  
 a A K G T L \* R L C R P V L \* N F K S G T -  
 b P K E P F R D Y V D R F F K T L R A E Q -  
 c Q R N P L E T M \* T G S L K L \* E R N K -  
  
 AGCTACGCAAGGTGAAAAATTGGATGACAGACACCTTGGTGGTCCAAATGCGAACCC  
 1081 -----+-----+-----+-----+-----+-----+-----+-----+-----+-----+ 1140  
 TCGATGCGTCCACATTAAACCTACTGTCTGTGGAAACAACCGAGTTTACGCTTGGG  
 a S Y A R C K K L D D R H L L V G P K C E P -  
 b A T Q G V K N W M T D T L L V Q N A N P -  
 c L R K V \* K I G \* Q T P C W S K M R T Q -

**Fig. 8/c**

AGATTGTAAGACCATTTAAAGAGCATTAGGACCAGGGGCTCAATAGAAGAAAATGATGAC  
 1141 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 TCTAACATTCTGTAAAATTCTCGTAATCCTGGTCCCGAAGTTATCTCTTTACTACTG  
 a R L \* D H F K S I R T R G F N R R N D D -  
 b D C K T I L R A L G P G A S I E E M M T -  
 c I V R P F \* E H \* D Q G L Q \* K K \* \* Q -  
 1200  
 AGCATGTCAGGGAGTGGGAGGACCTAGCCATAAAAGCAAAGTGTGGCCGAGGCAATGAG  
 1201 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 TCGTACAGTCCTCACCCTCTGGATCGGTATTTCGTTTACAACCGGCTCCGTTACTC  
 a S M S G S G R T \* P \* S K S V G R G N E -  
 b A C Q G V G G P S H K A K V L A E A M S -  
 c H V R E W E D L A I K Q K C W P R Q \* A -  
 1260  
 CCAACAAACAGTGCCATACTGATGCAGAGAAGCAATTAAAGGCTCTAAAAGAATTGT  
 1261 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 GGTTTGTTGTCACGGTATGACTACGTCTTCGTTAAATTCGAGATTTCCTAACCA  
 a P N K Q C H T D A E K Q F \* R L \* K N C -  
 b Q T N S A I L M Q R S N F K G S K R I V -  
 c K Q T V P Y \* C R E A I L K A L K E L L -  
 TAAATGTTCAACTGTGGCAAGGAAGGGCACATAGCCAGAAATTGCAAGGGCCCTAGGAA  
 1320  
 1321 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 ATTTACAAAGTTGACACCGTCTTCCTGGTATCGTCTTAACTGCTCCGGGATCCTT  
 a \* M F Q L W Q G R A H S Q K L Q G P \* E -  
 b K C F N C G K E G H I A R N C R A P R K -  
 c N V S T V A R K G T \* P E I A G P L G K -  
 1380  
 AAAGGGCTGTGGAAATGTGGAAAAGAAGGACACCAAATGAAAGATTGACTGAGAGACA  
 1381 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 TTTCCGACAACCTTACACCTTTCTTCCTGTGGTTACTTCTAACATGACTCTCTGT  
 a K G L L E M W K R R T P N E R L Y \* E T -  
 b K G C W K C G K E G H Q M K D C T E R Q -  
 c R A V G N V E K K D T K \* K I V L R D R -  
 1440  
 GGCAATTTCAGGGAAATCTGGCCCTCCCACAAGGGAGGGCCAGGGAAATTCTTC  
 1441 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 CCGGTTAAAAAAATCCCTTTAGACCGGGAGGGTGTCCCTCCGGTCCCTAAAAGAAGT  
 a G Q F F R E N L A L P Q G R A R E F S S -  
 b A N F L G K I W P S H K G G P G N F L Q -  
 c P I F \* G K S G P P T R E G Q G I F F R -  
 1500  
 GAACAGACCAAGGCCAACAGCCCCACCAGAGGAGAGCTCAGGTTGgGGAAGAGACAAC  
 1501 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 CTTGCTGGTCTCGGTTGTCGGGGGGTCTCTCGAAGTCCAACCCCTCTCTGTTG  
 a E Q T R A N S P T R G E L Q V W G R D N -  
 b N R P E P T A P P E E S F R F G E E T T -  
 c T D Q S Q Q P H Q R R A S G L G K R Q Q -  
 1560  
 AACTCCATCTCAGAAGCAGGAGCCAATAGACAAGGAACATATCTTAACTTCCCTCAA  
 1561 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 TTGAGGTAGAGTCTTCGTCCTCGGTTATCTGTTCTGATATAGGAATTGAAGGGAGTT  
 a N S I S E A G A N R Q G T I S F N F P Q -  
 b T P S Q K Q E P I D K E L Y P L T S L K -  
 c L H L R S R S Q \* T R N Y I L \* L P S N -  
 1620  
 ATCACTTTGGCAACGACCCCTCGTCACAATAAGATAGGGGGCAATTAAAGGAAGCT  
 1621 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 TAGTGAGAAACCGGTTGCTGGGGAGCAGTGTATTCTATCCCCCGTTAATTCTCGA  
 a I T L W Q R P L V T I K I G G Q L K E A -  
 b S L F G N D P S S Q \* R \* G G N \* R K L -  
 c H S L A T T P R H N K D R G A I K G S S -  
 1680  
 CTATTAGATACAGGAGCAGGTGATACAGTATTAGAAGACCTGAATTGCCAGGGAAATGG  
 1681 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 GATAATCTATGTCCTCGTCACATGTCATAATCTCTGGACTTAAACGGTCCCTTAC  
 a L L D T G A G D T V L E D L N L P G K W -  
 b Y \* I Q E Q V I Q Y \* K T \* I C Q G N G -  
 c I R Y R S R \* Y S I R R P E F A R E M E -

Fig. 8/d

**Fig. 8/e**

TATCAGTACAATGTA CTTCCACAGGGATGGAAAGGATCACTAGCAATATTCAAAGTAGC  
 2341 +-----+-----+-----+-----+-----+-----+-----+-----+  
 ATAGTCATGTTACATGAAGGTGTCCTACCTTCTAGTGATCGTTATAAGGTTCATCG  
 Y Q Y N V L P Q G W K G S L A I F Q S S -  
 I S T M Y F H R D G K D H \* Q Y S K V A -  
 S V Q C T S T G M E R I T S N I P K \* H -  
 ATGACAAAAACCTTAGAGCCTTTAGAAAACAAAATCCAGGCATA GTTATCTATCAATAC  
 2401 +-----+-----+-----+-----+-----+-----+-----+-----+  
 TACTGTTTTGGAATCTCGAAAATCTTGTGTTAGGTCCGTATCAATAGATAGTTATG  
 M T K T L E P F R K Q N P G I V I Y Q Y -  
 \* Q K P \* S L L E N K I Q A \* L S I N T -  
 D K N L R A F \* K T K S R H S Y L S I H -  
 ATGGATGATTGTATGAGGATCTGACTTAGAGATAGGGCAGCATAGAACAAAAATAGAG  
 2461 +-----+-----+-----+-----+-----+-----+-----+-----+  
 TACCTACTAACATACATCCTAGACTGAATCTCTATCCCGTCGTATCTGTGTTATCTC  
 M D D L Y V G S D L E I G Q H R T K I E -  
 W M I C M \* D L T \* R \* G S I E Q K \* R -  
 G \* F V C R I \* L R D R A A \* N K N R G -  
 GAACTGAGACAAACATTGTTGAGGTGGGATTTACACACCCAGACAAAGAACATTAGAAA  
 2521 +-----+-----+-----+-----+-----+-----+-----+-----+  
 CTTGACTCTGTTGAAACAACTCCACCCCCAAATGGTGTGGCTGTTCTTGTAAATCTT  
 E L R Q H L L R W G F T T P D K K H \* K -  
 N \* D N I C \* G G D L P H Q T R N I R K -  
 T E T T F V E V G I Y H T R Q E T L E R -  
 GAACCTCCATTCTTGATGGGTATGAACCTCCATCCTGACAAATGGACAGTACAGCCT  
 2581 +-----+-----+-----+-----+-----+-----+-----+-----+  
 CTTGGAGGTAAAGAAACCTACCCCATACTTGAGGTAGGACTGTTACCTGTATGTCGGA  
 E P P F L W M G Y E L H P D K W T V Q P -  
 N L H F F G W G M N S I L T N G Q Y S L -  
 T S I S L D G V \* T P S \* Q M D S T A Y -  
 ACACAGCTGCCAGAAAAGATAGCTGACTGTCAATGATATAACAAAAGTTAGGGAAAA  
 2641 +-----+-----+-----+-----+-----+-----+-----+-----+  
 TGTCTCGACGGCTTTTCTATCGACCTGACAGTTACTATATGTTTCAATCACCTTTT  
 T Q L P E K D S W T V N D I Q K L V G K -  
 H S C Q K K I A G L S M I Y K S \* W E N -  
 T A A R K R \* L D C Q \* Y T K V S G K I -  
 TTAAACTGGCAAGTCAGATTATCTGGAATTAAAGTAAGGCAACTTGTAACTCCTT  
 2701 +-----+-----+-----+-----+-----+-----+-----+-----+  
 AATTGACCCGTTCACTAAATAGGACCTTAAATTCATCCGTTGAAACATTGAGGAA  
 L N W A S Q I Y P G I K V R Q L C K L L -  
 \* T G Q V R F I L E L K \* G N F V N S L -  
 K L G K S D L S W N \* S K A T L \* T P \* -  
 AGGGGGCCAAAGCACTAACAGACATAGTACCAACTAATGAGAACAGAACATTAGAATTG  
 2761 +-----+-----+-----+-----+-----+-----+-----+-----+  
 TCCCCCGGTTCTGATTGTCATGATCTGGTATTGACTCTTCGCTCTTAATCTTAAC  
 R G A K A L T D I V P L T E E A E L E L -  
 G G P K H \* Q T \* Y H \* L K K Q N \* N W -  
 G G Q S T N R H S T T N \* R S R I R I G -  
 GCAGAAAACAGGGAAATTCTAAAAGAACCACTGAGGTATACTATGACCCATCAAAA  
 2821 +-----+-----+-----+-----+-----+-----+-----+-----+  
 CGTCTTTGTCCCTTAAAGATTCTGGTATGACCTCATATGATAGACTGGTAGTTT  
 A E N R E I L K E P V H G V Y Y D P S K -  
 Q K T G K F \* K N Q Y M E Y T M T H Q K -  
 R K Q G N S K R T S T W S I L \* P I K R -  
 GACTTGATAGCTGAAATACAGAACAGGGCAGGAACAATGGACATATCAAATTACAA  
 2881 +-----+-----+-----+-----+-----+-----+-----+-----+  
 CTGAACATCGACTTTATGTCCTTGTCCCCGTCCTTGTACCTGTATAGTTAAATGGTT  
 D L I A E I Q K Q G Q E Q W T Y Q I Y Q -  
 T \* L K Y R N R G R N G H I K F T K -  
 L D S \* N T E T G A G T M D I S N L P R -

Fig. 8/f

Fig. 8/g

**Fig. 8/h**

AATTTTAAAAGAAAAGGGGGATTGGGGGTACAGTCAGGGAAAGAATAATAGATATA  
 4141 -----+-----+-----+-----+-----+-----+-----+-----+ 4200  
 TTAAAATTTCTTCCCCCTAACCCCCCATGTCACGGCCCTTCTTATTATCTATAT  
 a N F K R K G G I G G Y S A G E R I I D I -  
 b I L K E K G G L G G G T V Q G K E \* \* I \* -  
 c F \* K K R G D W G V Q C R G K N N R Y N -

ATAGCAACAGACATACAAACTAAAGAATTACAAAAACAGATTACAAAAATTCAAATTTT  
 4201 -----+-----+-----+-----+-----+-----+-----+-----+ 4260  
 TATCGTTGTCGTATGTTGATTCTTAATGTTTGTCTAATGTTTAAAGTTTAAAAA  
 a I A T D I Q T K E L Q K Q I T K I Q N F -  
 b \* Q Q T Y K L K N Y K N R L Q K F K I F -  
 c S N R H T N \* R I T K T D Y K N S K F S -

CGGGTTTATTACAGAGACAGCAGAGACCCAGTTGGAAAGGACCAGCCAAACTACTCTGG  
 4261 -----+-----+-----+-----+-----+-----+-----+-----+ 4320  
 GCCCAAATAATGTCCTGTCGTCTGGGTCAACCTTCTGGTGGTGGATGAGACC  
 a R V Y Y R D S R D P S W K G P A K L L W -  
 b G F I T E T A E T P V G K D Q P N Y S G -  
 c G L L Q R Q Q R P Q L E R T S Q T T L E -

AAAGGTGAAGGGGCAGTAATAATACAAGATAATAGTACATAAAGTAGTACCAAGGAGG  
 4321 -----+-----+-----+-----+-----+-----+-----+-----+ 4380  
 TTTCCACTTCCCGTCATTATTGTTCTATTACTGTTGTTCCATCATGGTCTCC  
 a K G E G A V I I Q D N S D I K V V P R R -  
 b K V K G Q \* \* Y K I I V T \* R \* Y Q G G -  
 c R \* R G S N N T R \* \* \* H K G S T K E E -

AAAGCAAAATCATTAAGGACTATGGAAAACAGATGGCAGGTGCTGATTGTGTCAGGT  
 4381 -----+-----+-----+-----+-----+-----+-----+-----+ 4440  
 TTTCGTTTTAGTAATTCTGATACCTTTGTCTACCGTCCACGACTAACACACCGTCCA  
 a K A K I I K D Y G K Q M A G A D C V A G -  
 b K Q K S L R T M E N R W Q V L I V W Q V -  
 c S K N H \* G L W K T D G R C \* L C G R \* -

AGACAGGATGAAGATTAGAACATGGAATAGTTAGTAAACACCATATGTATGTTCAAG  
 4441 -----+-----+-----+-----+-----+-----+-----+-----+ 4500  
 TCTGTCCTACTCTAACTTGTACCTTATCAAATCATTGTTGTGTTACATAACAAAGTTC  
 a R Q D E D \* N M E \* F S K T P Y V C F K -  
 b D R M K I R T W N S L V K H H M Y V S R -  
 c T G \* R L E H G I V \* \* N T I C M F Q G -

GAGAGCTAATGGATGGTTTACAGACATCATTAGACAGACATCCAAAAGTAAGTTC  
 4501 -----+-----+-----+-----+-----+-----+-----+-----+ 4560  
 CTCTCGATTACCTACCAAAATGTCGTAGTAATACTGTCGTCTGTAGGTTTCAAG  
 a E S \* W M V L Q T S L \* Q Q T S K S K F -  
 b R A N G W F Y R H H Y D S R H P K V S S -  
 c E L M D G F T D I I M T A D I Q K \* V Q -

AGAAGTACACATCCCATTAGGAAAGGCTAAATTAGTAATAAAAACATATTGGGGTTGCA  
 4561 -----+-----+-----+-----+-----+-----+-----+-----+ 4620  
 TCTTCATGTGAGGGTAATCTTCCGATTTAATCATTATTTGTATAACCCCAACGT  
 a R S T H P I R K G \* I S N K N I L G V A -  
 b E V H I P L G K A K L V I K T Y W G L Q -  
 c K Y T S H \* E R L N \* \* \* K H I G G C R -

GACAGGAGAAAGAGATCGGCATTTGGTCATGGAGTCTCCATAGAATGGAGATTGAGAAG  
 4621 -----+-----+-----+-----+-----+-----+-----+-----+ 4680  
 CTGTCCCTTCTCTAGCGTAAACCCAGTACCTCAGGGTATCTTACCTCTAACTCTTC  
 a D R R K R S A F G S W S L H R M E I E K -  
 b T G E R D R H L G H G V S I E W R L R R -  
 c Q E K E I G I W V M E S P \* N G D \* E D -

ATATACCACAAATAGAACCTGGCCTGGCAGACCAAGCTAATTCAATTGTATTITGA  
 4681 -----+-----+-----+-----+-----+-----+-----+-----+ 4740  
 TATATGGTGTGTTATCTGGACGGACCGTCTGGTCGATTAAGTAAACATAATAAAACT  
 a I Y H T N R T W P G R P A N S F V L F \* -  
 b Y T T Q I E P G L A D Q L I H L Y Y F D -  
 c I P H K \* N L A W Q T S \* F I C I I L I -

Fig. 8/i

**Fig. 8/j**

GTGCTGCTTCATTGCCAAGTTGTTACGAAAAAGGCTTAGGCATCTTCTATGGCAG 5340  
 5281 -----+-----+-----+-----+-----+-----+  
 CACGACGAAAGTAACGGTCAAACAAAAGTCTTTCCGAATCGTAGAAGATAACCGTC  
 a V L L S L P S L F H E K R L R H L L W Q -  
 b C C F H C Q V C F T K K G L G I F Y G R -  
 c A A F I A K F V S R K K A \* A S S M A G -  
  
 GAAGAACGAAAGACAGCGACGAAGCGCTCATCGAAGCAGTGAGGATCATCAAAATCCTAT 5400  
 5341 -----+-----+-----+-----+-----+-----+  
 CTTCTTCGCTCTGCGCTTCGCGAGTAGCTTCGTCACTCCTAGTAGTTTAGGATA  
 a E E A K T A T K R S S K Q \* G S S K S Y -  
 b K K R R Q R R S A H R S S E D H Q N P I -  
 c R S E D S D E A L I E A V R I I K I L Y -  
  
 ATCAAAGCAGTAAGTAGTAAATGTAATGCAAGCTTAAACCATTAGCAATAGTAGCCTT 5460  
 5401 -----+-----+-----+-----+-----+-----+  
 TAGTTTCGTCAATTCACTATTACATTACCGTTGAAATTGGTAAATCGTTATCATCGGAA  
 a I K A V S S K C N A S F N H F S N S S L -  
 b S K Q \* V V N V M Q A L T I L A I V A L -  
 c Q S S K \* \* M \* C K L \* P F \* Q \* \* P \* -  
  
 AGTAGTAGCAACAATAATAGCAATAGTTGTTGGACCATAGTATTCAAGATAATAGGAA 5520  
 5461 -----+-----+-----+-----+-----+-----+  
 TCATCATCGTTGTTATTATCGTTATCAACACACCTGGTATCATAAGTATCTTATATCCTT  
 a S S S N N N S N S C V D H S I H R I \* E -  
 b V V A T I I A I V V W T I V F I E Y R K -  
 c \* \* Q Q \* \* Q \* L C G P \* Y S \* N I G K -  
  
 AATATTAAGACAGAAAAAAATAGACAGGTTAATTGATAGAATAAGAGAAAGAGCAGAAGA 5580  
 5521 -----+-----+-----+-----+-----+-----+  
 TTATAATTCTGCTTTTTTATCTGCTCAAATTAACTATCTTATTCTCTTCGTCCTCT  
 a N I K T E K N R Q V N \* \* N K R K S R R -  
 b I L R Q K K I D R L I D R I R E R A E D -  
 c Y \* D R K K \* T G \* L I E \* E K E Q K T -  
  
 CAGTGGCAATGAGGTGACGGGATCAGGAAGAATTATGGCATTATGGAGATGGGGCA 5640  
 5581 -----+-----+-----+-----+-----+-----+  
 GTCACCGTTACTCCCACTGCCCTAGTCCTCTTAATAGCGTAAATACCTCTACCCGT  
 a Q W Q \* G \* R G S G R I I G I Y G D G A -  
 b S G N E G D G D Q E E L S A F M E M G H -  
 c V A M R V T G I R K N Y R H L W R W G T -  
  
 CCATGCTCCTGGGATGTTGATGATCAGTAGTGTGCTAGGAAACTTGTGGGTACAGTCT 5700  
 5641 -----+-----+-----+-----+-----+-----+  
 GGTACGAGGAACCCCTACAACACTAGTCATCACGACATCCTTGAAACACCCAGTGTCA  
 a P C S L G C \* \* S V V L \* E T C G S Q S -  
 b H A P W D V D D Q \* C C R K L V G H S L -  
 c M L L G M L M I S S A V G N L W V T V Y -  
  
 ATTATGGGGTACCTGTATGGAAAGGGCAACCACCACTTATTGTGCTACAGATGCTA 5760  
 5701 -----+-----+-----+-----+-----+-----+  
 TAATACCCCCATGGACATACCTTCCCCGGTGGTGGTGAATAAAACACGTAGTCTACGAT  
 a I M G Y L Y G K G Q P P L Y F V H Q M L -  
 b L W G T C M E R G N H H F I L C I R C \* -  
 c Y G V P V W K G A T T T L F C A S D A K -  
  
 AAGCATATGATACAGAGGTACATAATGTTGGCTACACATGCCTGTACCCGACAGACC 5820  
 5761 -----+-----+-----+-----+-----+-----+  
 TTCTGTATACTATGTCTCCATGTATTACAAACCCGATGTACGGACACATGGGGCTCTGG  
 a K H M I Q R Y I M F G L H M P V Y P Q T -  
 b S I \* Y R G T \* C L G Y T C L C T R R P -  
 c A Y D T E V H N V W A T H A C V P A D P -

22/27

Fig. 8/k

CCAACCCACAAGAAATGGTTTGAAAATGTAACAGAAAATTAACTATGGAAAAATG  
 5821 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 GGTTGGGTGTTCTTACCAAAACCTTTACATTGCTTTAAAATTGTACACCTTTAC  
 P T H K K W F W K M \* Q K I L T C G K M -  
 b Q P T R N G F G K C N R K F \* H V E K \* -  
 c N P Q E M V L E N V T E N F N M W K N E -  
  
 AAATGGTAAATCAGATGCAGGAAGATGTAATCAGTTATGGGATCAAAGCTAAACCAT  
 5881 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 TTTACCATTTAGTCTACGCCCTCTACATTAGTCATAACCTAGTTGGATTTGGTA  
 a K W \* I R C R K M \* S V Y G I K A \* N H -  
 b N G K S D A G R C N Q F M G S K P K T M -  
 c M V N Q M Q E D V I S L W D Q S L K P C -  
 GTGTAAAGGTGACCCCCACTCTGTGTCACTTAGAATGTTAGAATGTTAGCACTAATGTA  
 5941 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 CACATTTCACTGGGTGAGACACAGTGAATCTTACATCTTACATCGTCATTATCAT  
 a V \* S \* P H S V S L \* N V E M L A V I V -  
 b C K V D P T L C H F R M \* K C \* Q \* \* \* -  
 c V K L T P L C V T L E C R N V S S N S N -  
  
 ATGATACCTACCATGAGACCTACCATGAGAGCATGAAGGAAATGAAAATTGCTTTCA  
 6001 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 TACTATGGATGGTACTCTGGATGGTACTCTCGTACTCTCTTTACTTTAACGAGAAAGT  
 a M I P T M R P T M R A \* R K \* K I A L S -  
 b \* Y L P \* D L P \* E H E G N E K L L F Q -  
 c D T Y H E T Y H E S M K E M K N C S F N -  
  
 ATGCAACCACAGTAGTAAGAGATAGGAAGCAGACAGTGTATGCACCTTTTATAGACTTG  
 6061 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 TACGGTGGTGTCACTTCTATCCTCGTCTGTACACATACGTGAAAAAAATATCTGAAC  
 a M Q P Q \* \* E I G S R Q C M H F F I D L -  
 b C N H S S K R \* E A D S V C T F L \* T \* -  
 c A T T V V R D R K Q T V Y A L F Y R L D -  
  
 ATATAGTACCACTACTAAGAAGAACTATAGTGAGAATTCTAGTAGTATTATAGATTAA  
 6121 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 TATATCATGGTGAATGATTCTTCTGATATCACTCTTAAGATCACTCATAATATCTAATT  
 a I \* Y H L L R R T I V R I L V S I I D \* -  
 b Y S T T Y \* E E L \* \* E F \* \* V L \* I N -  
 c I V P L T K K N Y S E N S S E Y Y R L I -  
  
 TAAATTGTAATACCTCAGCCATAACACAAAGCTGTCAAAGGTCACTTTGATCCAATTG  
 6181 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 ATTTAACATTATGGAGTCGGTATTGTTCTCGGACAGGGTTCCAGTGGAAAATAGGTTAAG  
 a \* I V I P Q P \* H K P V Q R S L L I Q F -  
 b K L \* Y L S H N T S L S K G H F \* S N S -  
 c N C N T S A I T Q A C P K V T F D P I P -  
  
 CTATACACTATTGCACTCCAGCTGGTTATGCAATTCTAAAGTGTAATGATAAGATATTCA  
 6241 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 GATATGTGAAACGTGAGGTGACCAATACGTTAAGATTTACATTACTATTCTATAAGT  
 a L Y T I A L O L V M Q F \* S V M I R Y S -  
 b Y T L L H S S W L C N S K V \* \* \* D I Q -  
 c I H Y C T P A G Y A I L K C N D K I F N -  
  
 ATGGGACAGGACCATGCCATAATGTTAGCACAGTACAATGTAACATGGGATTAAGCCAG  
 6301 +-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 TACCCCTGCTCTGGTACGGTATTACAATCGTGTCACTGTTACATGTTACCCCTAAATCGGTC  
 a M G Q D H A I M L A Q Y N V H M G L S Q -  
 b W D R T M P \* C \* H S T M Y T W D \* A S -  
 c G T G P C H N V S T V Q C T H G I K P V -

**Fig. 8/1**

TGGTATCAACTCAACTACTGTTAATGGTAGCCTAGCAGAAGGAGAATAATAATTAGAT  
 6361 +-----+-----+-----+-----+-----+-----+-----+ 6420  
 ACCATAGTGTAGTTGATGACAATTACCATCGGATCGTCCTCTTATTATTAATCTA  
 a W Y Q L N Y C \* M V A \* Q K E K \* \* L D -  
 b G I N S T T V K W \* P S R R R R N N N \* I -  
 c V S T Q L L L N G S L A E G E I I I I R S -

CTGAAAATCTGACAAACAATGTCAAACAAATAATAGTACATCTTAATCAATCTGTAGAAA  
 6421 +-----+-----+-----+-----+-----+-----+-----+ 6480  
 GACTTTTAGACTGTTTGTACAGTTGTTATTATCATGTAGAATTAGTTAGACATCTT  
 a L K I \* Q T M S K Q \* \* Y I L I N L \* K -  
 b \* K S D K Q C Q N N N S T S \* S I C R N -  
 c E N L T N N V K T I I V H L N Q S V E I -

TTGTATGTACAAGACCCGGAATAATACAAGAAAAAGTATAAGGATAGGACCAGGACAAA  
 6481 +-----+-----+-----+-----+-----+-----+-----+ 6540  
 AACATACATGTTCTGGGCCATTATGTTCTTCTCATATCCCTATCCTGGTCCTGTTT  
 a L Y V Q D P A I I Q E K V \* G \* D Q D K -  
 b C M Y K T R Q \* Y K K K Y K D R T R T N -  
 c V C T R P G N N T R K S I R I G P G Q T -

CATTCTATGCAACAGGAGACATAATAGGAGACATAAGACAAGCACATTGTAACATTAGTG  
 6541 +-----+-----+-----+-----+-----+-----+-----+ 6600  
 GTAAGATACTGTTCTCTGTATTATCCTCTGTATTCTCGTGTAACTTGTAAATCAC  
 a H S M Q Q E T \* \* E T \* D K H I V T L V -  
 b I L C N R R H N R R H K T S T L \* H \* \* -  
 c F Y A T G D I I G D I R Q A H C N I S E -

AAGATAAAATGGAATGAAACTTACAAAGGGTAAGTAAAAAATTAGCAGAACACTCCAGA  
 6601 +-----+-----+-----+-----+-----+-----+-----+ 6660  
 TTCTATTACCTTACCTTAACTGAAATGTTCCCATTCATTTTAATCGTCTGTGAAGGTCT  
 a K I N G M K L Y K G \* V K N \* Q N T S R -  
 b R \* M E \* N F T K G K \* K I S R T L P E -  
 c D K W N E T L Q R V S K K L A E H F Q N -

ATAAAACAATAAAATTGACATCATCCTCAGGAGGGACCTAGAAGTTACAACACATAGCT  
 6661 +-----+-----+-----+-----+-----+-----+-----+ 6720  
 TATTTGTTATTAAACCTGTTACCTTAACTGAAATGTTCCCATTCATTTTAATCGTCTGTGAAGGTCT  
 a I K Q \* N L H H P Q E G T \* K L Q H I A -  
 b \* N N K I C I I L R R G P R S Y N T \* L -  
 c K T I K F A S S S G G D L E V T T H S F -

TTAATTGTTAGGGAGAATTCTTCTATTGTAATACATCAGGCCGTGTTAATGGTCATACA  
 6721 +-----+-----+-----+-----+-----+-----+-----+ 6780  
 AATTAACATCCTCTTAAAGATAACATTATGTTAGTCGGACAAATTACACGTATGT  
 a L I V E E N F S I V I H Q A C L M V H T -  
 b \* L \* R R I F L L \* Y I R P V \* W C I H -  
 c N C R G E F F Y C N T S G L F N G A Y T -

CGCCTAATGGTACAAAAGTAATTCAAGCTCAATCATCACAACTCCATGCAGAATAAAGC  
 6781 +-----+-----+-----+-----+-----+-----+-----+ 6840  
 GCGGATTACCATGTTTCATTAAGTCGAGTTAGTAGTGTAGGGTACGTCTTATTTCG  
 a R L M V Q K V I Q A Q S S O S H A E \* S -  
 b A \* W Y K K \* F K L N H H N P M Q N K A -  
 c P N G T K S N S S S I I T I P C R I K Q -

AAATTATAAAATGTGGCAGGAGGTAGGACGAGCAATGTATGCCCTCCCATAAAAGGAA  
 6841 +-----+-----+-----+-----+-----+-----+-----+ 6900  
 TTTAATTTATACACCGTCTCCATCTGCTCGTACATACGGGAGGGTATTTCTT  
 a K L \* I C G R R \* D E Q C M P L P \* K E -  
 b N Y K Y V A G G R T S N V C P S H K R K -  
 c I I N M W Q E V G R A M Y A P P I K G N -

ACATAACATGTAATCAAATACACAGGACTACTATTGGTAGCTGATGGAGGAACAGAC  
 6901 +-----+-----+-----+-----+-----+-----+-----+ 6960  
 TGTTAGTGTACATTAGTTAGTGTCTGATGATAACCATGCACTACCTCCTGTCTCG  
 a T \* H V N Q I S Q D Y Y W Y V M E E Q S -  
 b H N M \* I K Y H R T T I G T \* W R N R A -  
 c I T C K S N I T G L L L V R D G G T E P -

**Fig. 8/m**

Fig. 8/n

**Fig. 8/o**

TGCAATAAAATGGGGGGCAAGTGGTCGAAAAGTAGCATAGTTGGATGCCCTGCTATAAGG  
 8161 +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 ACGTTATTACCCCGTCAACAGCTTACATCGTATCAACCTACCGGACGATATTCC 8220  
 a C N K M G G K W S K S S I V G W P A I R -  
 b A I K W G A S G R K V A \* L D G L L \* G -  
 c Q \* N G G Q V V E K \* H S W M A C Y K G -  
 GAGAGAATGAGAAGAACTGAGGCCAGCAGCAGATGGGGGGAGCAGTATCTCGAGACCTG  
 8221 +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 CTCTCTTACTCTTCTTGACTCGGTGCTCACCCACCCCTCGTCATAGAGCTCTGGAC 8280  
 a E R M R R T E P A A D G V G A V S R D L -  
 b R E \* E E L S Q Q Q M G W E Q Y L E T W -  
 c E N E K N \* A S S R W G G S S I S R P G -  
 GAAAAACATGGAGCAATCACGAGTAGCAATACAGCAGCTACTAATGAGGATTGTGCCTGG  
 8281 +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 CTTTTGTACCTCGTTAGTGCATCGTTATGTCGTCGATTAACCTCTAACACGGACC 8340  
 a E K H G A I T S S N T A A T N E D C A W -  
 b K N M E Q S R V A I Q Q L L M R I V P G -  
 c K T W S N H E \* Q Y S S Y \* \* G L C L A -  
 CTGGAAGCACAGAGGGGGGGAGGTGGGTTTCCAGTCAGACCTCAGGTACCTTAAGA  
 8341 +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 GACCTTCGTGTTCTCCCTCCCTCACCCAAAGGTCACTCTGGAGTCCATGGAAATTCT 8400  
 a L E A Q E E G E V G F P V R P Q V P L R -  
 b W K H K R R G R W V F Q S D L R Y L \* D -  
 c G S T R G G G G F S S Q T S G T F K T -  
 CCAATGACTTACAAGGGAGCTGTAGATCTTCTTTTAAAGAAAAGGGGGACTG  
 8401 +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 GGTTACTGAATGTTCCCTCGACATCTAGAATCGAAGAAAATTTCTTCCCCCTGAC 8460  
 a P M T Y K G A V D L S F F L K E K G G L -  
 b Q \* L T R E L \* I L A S F \* K K R G D W -  
 c N D L Q G S C R S \* L L F K R K G G T G -  
 GAAGGGTTAATTACTCTAAGAAAAGCAAGAGATCCTTGATTGTGGGTCTATCACACA  
 8461 +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 CTTCCAATTAATGAGATTCTTTCCGTTCTAGGAACTAAACACCCAGATAGTGTGT 8520  
 a E G L I Y S K K R Q E I L D L W V Y H T -  
 b K G \* F T L R K G K R S L I C G S I T H -  
 c R V N L L \* E K A R D P \* F V G L S H T -  
 CAAGGCTACTCCCTGATTGGCACACTACACACCAGGACCAAGGGTCAGATTCCACTG  
 8521 +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 GTTCGATGAAGGGACTAACCGTGTGATGTGTGGTCTGGTCCCTCAGTCTAAGGGTGAC 8580  
 a Q G Y F P D W H N Y T P G P G V R F P L -  
 b K A T S L I G T T T H Q D Q G S D S H \* -  
 c R L L P \* L A Q L H T R T R G Q I P T D -  
 ACTTTGGGTGGTCTTCAAGCTAGTACCGTTGACCAAGGGAGTAGAAGAGGCCAAC  
 8581 +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 TGAAAACCCACACGAAGTCGATCATGGTCAACTGGGTTCCCTCATCTTCTCCGGTTG 8640  
 a T F G W C F K L V P V D P R E V E E A N -  
 b L L G G A S S \* Y Q L T Q G K \* K R P T -  
 c F W V V L Q A S T S \* P K G S R R G Q R -  
 GAGGGAGAAGACAACGTGCTTGCTACACCCGTGTGCCAGCATGGAATGGAGGATGATCAC  
 8641 +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 CTCCTCTTCTGTTGACGAACGATGTGGGACACACGGTCGACCTACCTCTACTAGTG 8700  
 a E G E D N C L L H P V C Q H G M E D D H -  
 b R E K T T A C Y T L C A S M E W R M I T -  
 c G R R Q L L A T P C V P A W N G G \* S Q -  
 AGAGAAAGTATTAAGTGGAAAGTTGACACTCAACTAGCACACAGACACAGGGCCCGAA  
 8701 +-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+  
 TCTCTTCAATAATTACCTTCAAACGTGTCAGTTGATCGTGTGTCTGTGTGGCCGGCTT 8760  
 a R E V L K W K F D S Q L A H R H R A R E -  
 b E K Y \* S G S L T V N \* H T D T G P A N -  
 c R S I K V E V \* Q S T S T Q T Q G P R T -

**Fig. 8/p**

## SEQUENZPROTOKOLL

<110> Geneart GmbH  
5           <120> Das Genom des HIV-1 Intersubtyps (C/B') und seine  
             Anwendungen  
  
<130> WAG-001 PCT  
10          <140> xx  
             <141> 2000-11-16  
  
<150> DE 199 55 089.1  
15          <151> 1999-11-16  
  
<160> 3  
  
<170> PatentIn Ver. 2.1  
20          <210> 1  
             <211> 9078  
             <212> DNA  
             <213> Human immunodeficiency virus  
25          <400> 1  
             aatctctagc agtggcgccc gaacaggac ttgaaagcga aagtaagacc agaggagatc 60  
             tctcgacgca ggactcggt tgctgaagt cactcgccaa gaggcgagag cggcgactgg 120  
             tgagtacgcc aattatattt gactacgga ggctagaagg agagagatgg gtgcgagagc 180  
             gtcaatatta agaggggaa aattagataa atggaaaaaa attaggttaa ggccaggggg 240  
             aaagaaaacac tatatgctaa aacaccttgt atggcaagc agggagctgg aaagattgc 300  
             acttaaccct ggccttttag agacatcaga aggctgtaaa caaataatga aacagctaca 360  
             atcagctttt cagacaggaa cagaggaact tagatcatta ttcaacacag tagcaactcc 420  
             ctattgtgtc catacagaga tagatgtacg agacaccaga gaagccttag acaagataga 480  
             ggaagaacaa aacaaaattc agcaaaaaac acagcaggca aaggaggctg acgggaaggt 540  
             cagtcaaaat ttccttatag tacagaatct ccaaggcata 600  
             accttagaact ttatgtcat gggtaaaaat ggtagaagag aaggctttt gcccagaagt 660  
             aataccccatg ttttcagcgt tatcagaagg agccacccca caagattaa acaccatgt 720  
             aaacacagtg gggggacatc aagcagctat gcaaataat aagatacca tcaatgaaga 780  
             ggctgcagaa tggatagat tacatccagt acatgcaggc cctattgcac caggccaaat 840  
             gagagaacca aggggaagtg acatgcagg aactactgt aacctacagg aacaaatgc 900  
             atggatgacg agtaacccac ctgtttcagt aggagacatc tataaaatggataattct 960  
             gggattaaat aaaatgtatc gatatgtatc ccctaccatc attctggaca taaaacaagg 1020

gccaaaggaa cccttagag actatgtaga ccgttcttt aaaacttaa gagcggaca 1080  
agtcacgcaa ggtgtaaaaa attggatgac agacaccttg ttggccaaa atgcgaaccc 1140  
agattgtaaag accatTTAA gaggattttt accaggggct tcaatagaag aaatgtgac 1200  
agcatgtcg ggagtggag gacctagcca taaagaaaa gtgtggccg aggcaatgag 1260  
5 ccaaacaac agtgcatac tgatgcagag aagcaatttt aaaggctcta aaagaattgt 1320  
taaatgtttc aactgtggca aggaaggca catagccaga aattgcaggg cccctaggaa 1380  
aaaggctgt tgaaatgtg gaaaagaagg acaccaaatg aaagattgtt ctgagagaca 1440  
ggccaatttt ttagggaaaa tctggccctc ccacaaggga gggccaggga attttctca 1500  
gaacagacca gagccaacag cccccaccaga ggagagcttc aggtttgggg aagagacaac 1560  
10 aactccatct cagaaggcagg agccaataga caaggaacta tatccttaa cttccctcaa 1620  
atcactctt ggcaacgacc cctcgtcaca ataaagatag gggggcaatt aaaggaagct 1680  
ctattagata caggagcagg tgatacagta tttagaagacc tgaatttgcc agggaaatgg 1740  
aaaccaaaaa tgataggggg aattggaggt tttatcaaag taagacagta tgaacagata 1800  
cccatagaaa ttgcggaca caaagctata ggtacagtt tagtaggacc tacacctgtc 1860  
15 aacataattt gaagaaatct gttgacttag cttgggtgca ctttaaattt tccaatcagt 1920  
cccattgaaa ctgtaccagt aaaattaaag ccaggaatgg atggcccaa ggttaaaca 1980  
tggccattga cagaagagaa aataaaagca ttaacagcaa tttgtatgt aatggagaaa 2040  
gaaggaaaaa ttacaaaaat tggcctgaa aatccatata acactccat atttgccata 2100  
aaaaagaagg acagtactaa gtggagaaag ttagtagatt tcagggact caataaaaga 2160  
20 actcaagatt ttggaaagt tcaatttagga ataccacacc cagcagggtt aaaaaagaaa 2220  
aaatcgtga cagtactgga tggggggat gcatatTTTT caattccctt atatgaagac 2280  
ttcaggaagt atactgcatt caccatacct agtagaaaca atgaaacacc agggatttagg 2340  
tatcgtaca atgtacttcc acaggatgg aaaggatcac tagcaatatt ccaaagtgc 2400  
atgacaaaaaa ccttagagcc ttttagaaaa caaaatccag gcatagttat ctatcaatac 2460  
25 atggatgatt tggatgttggg atctgactta gagatagggc agcatagaac aaaaatagag 2520  
gaactgagac aacattttt gaggggggaa ttaccacac cagacaagaa acattagaaa 2580  
gaacccat ttctttggat gggatgaa ctccatcctg acaaattggc agtacagcct 2640  
acacagctgc cagaaaaaga tagctggact gtcaatgata tacaatggg aatggggaaa 2700  
ttaaactggg caagtcagat ttatcctgaa attaaatggg gcaactttt taaactccctt 2760  
30 agggggccca aagcaactaac agacatagta ccactaactg aagaaggcaga attagaattt 2820  
gcagaaaaaca gggaaattct aaaagaacca gtacatggag tataactatga cccatcaaaa 2880  
gacttgatag ctgaaataca gaaacagggg caggaacaat ggacatatac aatttaccaa 2940  
gaaccattca aaaatctaaa aacagggaaatg tggcaaaaaa tgaggactgc ccacactaat 3000  
gatgtaaaac aattaacaga ggctgtcgaaaatggca tggaggat agtaatatgg 3060  
35 gggaaaaactc ctaaattttt attaccatc caaaaagaaa catggagac atggggaca 3120  
gacttggc aagccatcg gattcctgat tggggatTTT ttaataccctt tcccttagta 3180  
aaattatggt accagctgaa aaaagatccc atagtagggat tagaaatctt ctatgtat 3240  
ggagcagcta atagggagac taaaatggaa aaagcagggt atgttactga cagagggagg 3300  
aagaaaaattt tttctctaac tggaaacaaca aatcagaaga ctgaattgca agcaatttgt 3360  
40 atagcttgc aagattcagg atcagaagta aacatagtaa cagattcaca gtatgcatta 3420  
gggatcattc aagcacaacc agataagagt gaatcagagt tagttaacca aataatagaa 3480  
caattaatga aaaaggaaag agtctacctg tcatgggtac cagcacataa aggaattgg 3540  
ggaaatgaac aagttagataa attagtaagt agtggaatca ggaaagtgc atttcttagat 3600

ggaatagata aagctcaaga agagcatgaa aagtatcaca gcaattggag agcaatggct 3660  
agtacttta atctgccacc catagtagca aaagaaaatag tggctagctg tgatcaatgt 3720  
cagctaaaag gagaagccat gcatggacaa gtagactgta gtccaggat atggcaatta 3780  
gattgtacac atttagaagg aaaaatcatc ctggtagcag tccatgtac cagtggctac 3840  
5 atggaagcag aggttatccc agcagaaaca ggacaagaga cagcatactt tatactaaaa 3900  
ttagcagggaa gatggccagt caaagtaata catacagata atggtagtaa tttcaccagt 3960  
actgcagtttta aggcagcctg ttgggtggca ggtatccaac aggaatttgg aattccctac 4020  
agttccccaaa gtcagggaggt agttagaagcc atgaataaaag aattaaagaa aattataggg 4080  
caggtaagag atcaagctga gcacctaag acagcagtac taatggcagt attcattcac 4140  
10 aattttaaaa gaaaaggggg gattgggggg tacagtgcag gggaaagaat aatagatata 4200  
atagcaacag acatacaaac taaagaatta caaaaacaga ttacaaaaat tcaaaatttt 4260  
cggtttattt acagagacag cagagacccc agttggaaag gaccagccaa actactctgg 4320  
aaaggtgaag gggcagtaat aatacaagat aatagtgaca taaaggtat accaaggagg 4380  
aaagcaaaaa tcattaagga ctatggaaaa cagatggcag gtgctgatgg tggcaggt 4440  
15 agacaggatg aagatttagaa catggaatag ttttagaaaa caccatatgt atgtttcaag 4500  
gagagctaat ggatggttt acagacatca ttatgacagc agacatccaa aagtaagttc 4560  
agaagtacac atcccattag gaaaggctaa attagtaata aaaacatatt gggggttgca 4620  
gacaggagaa agagatcggc atttggtca tggagtctcc atagaatggg gattgagaag 4680  
atataccaca caaatagaac ctggcctggc agaccagctt attcattttt attatttga 4740  
20 ttgttttgcgactctgata taaggaaagc catatttaga cacatagttt ttccttaggt 4800  
tgactatcaa gcaggacata ataataaggt aggatctcta caataacttgg cactgacagc 4860  
attgataaaaa ccaaaaaaaa taaagccacc tctgcctagt atcaagaaat tagtagagga 4920  
tagatggAAC aatccccagg agatcgggg ccgcagaggg aaccacacaa tgaatggaca 4980  
ctagagcttc tagaggagct caagcaggaa gctgttagac actttcttag accatggctt 5040  
25 catagcttag gacaacatata ctatggaaaca tatgggata cttggcagg agtggaaagcc 5100  
ataataagaa ttctgcaaca actgctgtt attcatttca gaattgggtt tcagcatagc 5160  
agaataggca tttttagaca gagaagaaca agaaatggag ccagtaaatc ataaattaga 5220  
gccttggag catccaggaa gtcagctaa gactgctgtt aacagttgtt attgtaaaaa 5280  
gtgctgctttt cattgccaag tttgttccac gaaaaaaaggc tttaggcattt tctatggcag 5340  
30 gaagaagcga agacagcgcac gaagcgtca tcgaagcagt gaggatcatc aaaatctat 5400  
atcaaagcg taagtagtaa atgtaatgca agctttaacc atttttagcaa tagtagcctt 5460  
agtagtagca acaataatag caatagttgt gtggaccata gtattcatag aatataggaa 5520  
aatattaaga cagaaaaaaa tagacaggtt aattgtataga ataagagaaa gagcagaaga 5580  
cagttggcaat gagggtgacg gggatcagga agaattatcg gcatttatgg agatggggca 5640  
35 ccatgctctt tgggatgtt atgatcagta gtgctgttagg aaacctgtgg gtcacagttt 5700  
attatgggtt acctgtatgg aaaggggcaaa ccaccactttt attttgtgca tcagatgcta 5760  
aagcatatga tacagaggta cataatgtttt gggctacaca tgcctgtta cccgcagacc 5820  
ccaacccaca agaaatgggtt ttggaaaaatg taacagaaaaa tttaacatg tggaaaaatg 5880  
aaatggtaaa tcagatgcag gaagatgtaa tcagttttagt ggatcaaagc ctaaaaccat 5940  
40 gtgtaaagtt gacccactc tgggtcaactt tagaatgttag aatgttagc agtaatagta 6000  
atgataccta ccatgagacc taccatgaga gcatgaagga aatgaaaaat tgcactttca 6060  
atgcaaccac agtagtaaga gataggaagc agacagtgtta tgcactttt tataacttg 6120  
atatagtacc acttactaag aagaactata gtgagaattc tagtgagttat tataactttaa 6180

taaattgtaa tacctcagcc ataacacaag cctgtccaaa ggtcaacttt gatccaattc 6240  
ctatacacta ttgcaactcca gctggttatg caattctaaa gtgtaatgat aagatattca 6300  
atgggacagg accatgccat aatgttagca cagtacaatg tacacatggg attaagccag 6360  
tggtatcaac tcaactactg ttaaatggta gcctagcaga aggagaaaata ataattagat 6420  
5 ctgaaaatct gacaaacaat gtcaaaacaa taatagtaca tcttaatcaa tctgtagaaa 6480  
ttgtatgtac aagaccgcg aataatacaa gaaaaagtat aaggatagga ccaggacaaa 6540  
cattctatgc aacaggagac ataataggag acataagaca agcacattgt aacattagt 6600  
aagataaaatg gaatgaaact ttacaaaggg taagtaaaaa attagcagaa cactccaga 6660  
ataaaaacaat aaaatttgca tcatcctcag gaggggacct agaagttaca acacatagct 6720  
10 ttaattgtag aggagaattt ttctattgtat atacatcagg cctgtttat ggtgcataca 6780  
cgccataatgg tacaaaaagt aattcaagct caatcatcac aatcccattgc agaataaaagc 6840  
aaattataaa tatgtggcag gaggttaggac gagcaatgta tgcccttc 6900  
acataacatg taaatcaa atcacaggac tactattgtt acgtgtatgga ggaacagagc 6960  
caaatgatac agagacattc agacctggag gaggagatat gaggaacaat tggagaagt 7020  
15 aattatataa atataaagtg gtagaaatta agccattggg agtagcaccc actacaacaa 7080  
aaaggagagt ggtggagaga gaaaaaagag cagtgaaat aggagctgtg ttcccttgggt 7140  
tcttaggagt agcaggaagc actatggcg cgccgtcaat aacgctgacg gtacaggcca 7200  
gacaattgt gtctggata gtgcaacagc aaagcaattt gctgagggt atagaagcgc 7260  
aacagcatct gttgcaactc acggcttggg gcattaagca gctccagaca agagtccctgg 7320  
20 ctatagaaag atacctaagat gatcaacagc tccttagggat ttggggctgc tctggaaaac 7380  
tcatctgcac tactgctgta ccttggaaact ccagttggag taacaaatct caaaaagaga 7440  
tttggataa catgaccctgg atgcaatggg ataaagaaaat tagtaattac acaaacacag 7500  
tatacagggtt gcttgaagaa tcgaaaacc agcaggaaag gaatgaaaaa gatctattag 7560  
cattggacag ttggaaaaat ctatggagtt ggtttgacat aacaaattgg ctgtgtata 7620  
25 taaaaatatt cataataata gtaggaggct tgataggttt aagaataatt tttgctgtgc 7680  
tctctatagt aaatagagtt aggcaggat actcacctt gtcgtttcag acccttaccc 7740  
cgaacccagg gggacccgac aggctcgaa gaatcgaaga agaagggtgga aagcaagaca 7800  
gggacagatc cattcgatata gtgaacggat tcttagcgt tgcctggac gacctgcgga 7860  
acctgtgcct cttcagctac caccgattga gggacttcac attagtggca gcgagggtgg 7920  
30 tggaaacttct gggacgcaat agtctcagg gactacagag agggtgtggaa gcccttaat 7980  
atctggaaag tcttgcac tactgggtc aggagctaa aaagagtact attagtctgg 8040  
ttgataccat agcaatagca gtagctgaag gaacagatag gattatagaa ttagtacaag 8100  
gactttgtag agctatctac agcataccta gaagaataag acagggttt gaagcagctt 8160  
tgcaataaaa tggggggcaa gtggtcgaaa agtagcatag ttggatggcc tgctataagg 8220  
35 gagagaatgaa gaagaactga gccagcagca gatgggggtgg gaggcgtatc tcgagacctg 8280  
gaaaaacatg gagcaatcac gagtagcaat acagcagcta ctaatgagga ttgtgcctgg 8340  
ctggaaagcac aagaggagg ggaggtgggt ttccctgtca gacctcaggt acctttaaga 8400  
ccaatgactt acaagggagc tgttagatctt agcttctttt taaaagaaaa gggggactg 8460  
gaagggttaa ttactctaa gaaaaggcaa gagatccttg atttgtgggt ctatcacaca 8520  
40 caaggctact tccctgattt gcacaactac acaccaggac caggggtcag attccactg 8580  
acttttgggt ggtgcttcaa gctagtagca gttgacccaa gggaaagtaga agaggccaa 8640  
gagggagaag acaactgctt gctacaccct gtgtgccagc atggaatgga ggatgatcac 8700  
agagaagtat taaagtggaa gtttgacagt caactagcac acagacacag ggcccgca 8760

ctacatccgg agttttacaa agactgctga cacagaaggg actttccgcg gggacttcc 8820  
actggggcgt tcttaggaggt gtggcttgcg gggactggga gtggtaacc ctcaaatgct 8880  
gcatataagc agctgcttt cgcctgtact gggctctct agtcagacca gatctgagcc 8940  
tgggagctct ctggctaact agggAACCA ctgtttaagc ctcaataaag cttgccttga 9000  
5 ggggctagag cggccgcac cgcgggtggag ctccagctt tttcccttt agtgagggtt 9060  
aattgcgcgc tggcgatc 9078

<210> 2  
10 <211> 4288  
<212> DNA  
<213> Human immunodeficiency virus

<400> 2  
15 gcttagtgtacc taatgggcgc cagggccagc atcctgaggg gcggcaagct ggacaagtgg 60  
gagaagatca ggctgaggcc cggcgccaa aagcaactaca tgctgaagca cctgggtgtgg 120  
gccagcaggag agctggagag gttcgccctg aaccccccggcc tgctggagac cagcgaggcc 180  
tgaagcaga tcatgaagca gctgcagagc gcctgcaga cggcacccga ggagctgagg 240  
agcctgttca acaccgtggc cacccttac tgcgtgcaca ccgagatcga cgtgaggggac 300  
20 accagggagg ccctggacaa gatcgaggag gagcagaaca agatccagca gaagaccac 360  
caggccaagg aggccgacgg caaggtgagc cagaactacc ccatcggtca gaacctgcag 420  
ggccagatgg tgcaccagcc catcagcccc aggaccttga atgcattgggt gaaggtgggt 480  
gaggagaagg cttcagccc cgaggtgtatccatgcgttca gcgccttgcg cggggcgcc 540  
accccccagg acctgaacac catgctgaac accgtgggcg gccaccaggc cgccatgcag 600  
25 atcctgaagg acaccatcaa cgaggaggcc gcccggatggg acaggctgca ccccggtgcac 660  
gccccccca tcgccccccgg ccagatgagg gagcccaggc gcagcgacat cggccgcacc 720  
accagcaacc tgcaggagca gatcgccctgg atgaccagca acccccccgt gcccgtggc 780  
gacatctaca agagggttat catcctgggt taaaacaaga tcgtgaggat gtacagcccc 840  
accagcatcc tggacatcaa gcagggcccc aaggagccct tcaggacta cgtcgacagg 900  
30 ttcttcaaga ccctgagggc ggagcaggcc acccaggccg tgaagaactg gatgaccgac 960  
accctgttgc tgcagaacgc caaccggac tgcagaacca tcctgaggcc cctggggcccc 1020  
ggcccgacca tcgaggagat gatgaccgac tgccaggccg tggccggcccc cagccacaag 1080  
gccaagggtgc tggccggaggc catgagccag accaacagccg ccacccctgtat gcagaggagc 1140  
aacttcaagg gcagcaagag gatcgtaag tgcttcaact gggcaaggaa gggccacatc 1200  
35 gccaggaact gcagggcccc caggaagaag ggctgcttggaa agtgcggcaaa ggagggccac 1260  
cagatgaagg actgcaccgc gaggcaggcc aaccttctgg gcaagatctg gcccagccac 1320  
aaggggggcc cggcaactt cctgcagaac aggccccggc ccacccggcccc ccccgaggag 1380  
agcttcaggat tcgaggagga gaccaccacc cccagccaga agcaggagcc catcgacaag 1440  
gagctgtacc ccctgaccacg cctgaagagc ctgttccggca acgacccttccag cagccaggaa 1500  
40 ttcttcagggg agaacctggc cctgccccccag ggcaggccca gggagttcag cagcgaggc 1560  
accaggccca acagccccac cagggggcag ctgcagggtgt gggggcaggaa caacaacagc 1620  
atcagcgagg cggcgccaa caggcaggcc accatcgatc tcaacttccc ccagatcacc 1680  
ctgtggcaga gggccctgggt gaccatcaag atcggcgcc acgtgaagga gggccctgctg 1740

aacaccggcg ccggcgacac cgtgctggag gacctgaacc tgcccgaa gtggaagccc 1800  
aagatgatcg gcggcatcg cggttcata aaggtgaggc agtacgagca gatccccatc 1860  
gagatctgcg gccacaaggc catggcacc gtgctggtgg gccccacccc cgtgaacatc 1920  
atccggcagga acctgctgac ccagctgggc tgcacccctga acttcccat cagccccatc 1980  
5 gagaccgtgc ccgtgaagct gaagccggc atggacggcc ccaaggtgaa gcagtggccc 2040  
ctgaccggagg agaagatcaa ggccctgacc gccatctgcg acgagatgga gaaggaggc 2100  
aagatcacca agatcgcccc cgagaacccc tacaacaccc ccatcttcgc catcaagaag 2160  
aaggacagca ccaagtggag gaagctggt gacttcagg agctgaacaa gaggacccag 2220  
gacttctggg aggtgcagct gggcatcccc caccccgccg gcctgaagaa gaagaagac 2280  
10 gtgaccgtgc tggacgtggg cgacgcttac tttagcatcc ccctgtacga ggacttcagg 2340  
aagtacaccg ctttcacccat ccccgagg aacaacgaga ccccccggcat cagctaccag 2400  
tacaacgtgc tgccccaggg ctggaaggc agcctggca tcttccagag cagcatgacc 2460  
atcgaggagc ttagtctacag caagaagagg caggagatcc tggacctgt ggtgtaccac 2520  
acccagggct acttccctga ctggcacaac tacacccctg gccccggcgt gaggttcccc 2580  
15 ctgaccctcg gctgggtgctt caagctggtg cccgtggacc ccagggaggt ggaggaggcc 2640  
aacgagggcg aggacaactg cctgctgcac cccgtgtgcc agcacggcat ggaggacgac 2700  
cacagggagg tgctgaagtg gaagttcgac agccagctgg cccacaggca cagggccagg 2760  
gagctgcacc ccgagttcta caaggactgc atggcggca agtggagcaa gagcagcatc 2820  
gtgggctggc ccgcattcag ggagaggatg aggaggaccg agcccgccgc cgacggcgtg 2880  
20 ggcgcctgtga gcagggaccc ggagaagcac ggcgcctatca ccagcagcaa caccggcc 2940  
accaacgagg actgcgcctg gctggaggcc caggaggagg gcgaggtggg cttccctgt 3000  
agccccagg tgccccctgag gcccattgacc tacaaggccg ccgtggaccc gagcttcttc 3060  
ctgaaggaga agggcggcct ggagggcctg aggacgacc tggctggatgg gggcttcacc 3120  
accccgaca agaagcacca gaaggagccc cccttcgtt ggatgggcta cgagctgcac 3180  
25 cccgacaagt ggaccgtgca gcccacccag ctgcccggaga aggacagctg gaccgtgaac 3240  
gacatccaga agctgggtgg caagctgaac tggccagcc agatctaccc cggcatcaag 3300  
gtgaggcagc tgtgcaagct gctggggcgc gccaaggccc tgaccgacat cgtccccctg 3360  
accgaggagg ccgagctgga gctggccgag aacaggaga tcctgaagga gcccgtgcac 3420  
ggcgtgtact acgacccctag caaggacccctg atcggcggaga tccagaagca gggccaggag 3480  
30 cagtggaccc accagatcta ccaggagccc ttcaagaacc tgaagaccgg caagtacgcc 3540  
aagatgagga cggcccacac caacgacgtg aagcagctga ccgaggccgt gcagaagatc 3600  
gccatggagg gcatcgat ctggggcaag acccccaagt tcaggctgcc catccagaag 3660  
gagacctggg agacctggtg gaccgactac tggcaggcca cctggatccc cgagtggag 3720  
ttcgtgaaca cccctccctt ggtgaagctg tggatcage tggagaagga ccccatcggt 3780  
35 ggcgtggaga cttctacgt ggacggcc gccaacaggag agaccaagat cggcaaggcc 3840  
ggctacgtga ccgacagggg caggaagaaatcgtgagcc tgaccgagac caccaaccag 3900  
aagaccgagc tgcaggccat ctgcattgccttgcaggaca gccggcagcga ggtgaacatc 3960  
gtgaccgaca gccagtacgc cctggccatc atccaggccc agcccgacaa gagcggagac 4020  
gagctggtga accagatcat cgacgactg atgaagaagg agagggtgtt cctgagctgg 4080  
40 gtgccccccc acaagggcat cggcgacac gagcagggtgg acaagctggt gaggcaggcc 4140  
atcaggaagg tggatcgtggac cctggggcccttccaggaaatcgtgatc 4200  
taccaggata tggacgaccc tggatcgtggc agcggacccatgg agatcggccca gcacaggacc 4260  
aagtaaagat ctctcgagga gctcaagc 4288

<210> 3  
<211> 2605  
5 <212> DNA  
<213> Human immunodeficiency virus

<400> 3

10 gccggcggta ccgaattcgc cgccagcatg gacagggcca agctgctgct gctgctgctg 60  
ctgctgctgc tgccccaggc ccagggcgtg gcaacacctgt gggtgaccgt gtactacggc 120  
gtgcccgtgt ggaaggggcgc caccaccacc ctgttctgcg ccagcgacgc caaggctac 180  
gacaccgagg tgcacaacgt gtggggccacc cacgcctgcg tgcccgccga ccccaacccc 240  
caggagatgg tgctggagaa cgtgaccgag aacttcaaca tgtggaagaa cgagatggtg 300  
aaccagatgc aggaggacgt catcagcctg tgggaccaga gcctgaagcc ctgcgtgaag 360

15 ctgacccccc tctgcgtgac cctggagtgc aggaacgtga gcagcaacag caacgcacacc 420  
taccacgaga cctaccacga gagcatgaag gagatgaaga actgcagctt caacgcacacc 480  
accgtggtga gggacaggaa gcagaccgtg tacgcctgt tctacaggct ggacatcgtg 540  
cccccgtacca agaagaacta cagcgagaac agcagcgagt actacaggct gatcaactgc 600  
aacaccacgc ccatcaccca ggcctgcccc aaggtgaccc tcgaccccat ccccatccac 660

20 tactgcaccc ccggcggcta cgccatcctg aagtgcacg acaagatctt caacggcacc 720  
ggccctgcc acaacgtgag caccgtgcag tgcacccacg gcatcaagcc cgtggtgagc 780  
accgcgtgc tgctgaacgg cagcctggcc gagggcgaga tcatcatcag gagcgagaac 840  
ctgaccaaca acgtaaaaac catcatcgtg cacctgaacc agagcgtgga gatcgtgtc 900  
accaggcccc gcaacaacac caggaagagc atcaggatcg gccccggcca gacccatctac 960

25 gccaccggcg acatcatcgg cgacatcagg cagggccact gcaacatcag cgaggacaag 1020  
tggaacgaga ccctgcagag ggtgagcaag aagttgccc agcacttcca gaacaagacc 1080  
atcaagttcg ccagcagcag cggcggcgac ctggaggtga ccacccacag cttcaactgc 1140  
agggcgagc tcttctactg caacaccacg ggcctgttca acggcgccta caccggcaac 1200  
ggcaccaaga gcaacagcag cagcatcatc accatccct gcaggatcaa gcagatcatc 1260

30 aacatgtggc aggagggtgg cagggccatg tacgccttc ccatcaaggg caacatcacc 1320  
tgcaagagca acatcacccgg cctgctgctg gtgagggacg gggcaccgc gcccacgc 1380  
accgagaccc tcaggccccg cggcggcgac atgaggaaca actggaggag cgagctgtac 1440  
aagtacaagg tggtggagat caagccctg ggcgtggccc ccaccaccc caagaggagg 1500  
gtggtggaga gggagaagag ggcctgggc atcggcggc tttctggc cttctggc 1560

35 gtggccggca gcaccatggg cggccgcacg atcaccctga ccgtgcaggc caggcagctg 1620  
ctgagcggca tcgtgcagca gcagagcaac ctgctgaggc ccatcgaggc ccagcagcac 1680  
ctgctgcagc tgaccgtgt gggcatcaag cagctgcaga ccagggtgt ggcctcgag 1740  
agttacctga aggaccagca gctgctggc atctggggct gcaaggccaa gctgatctgc 1800  
accaccgcg tgccctggaa cagcagctgg agcaacaaga gccaaggaa gatctggac 1860

40 aacatgaccc ggtgcagtg ggacaaggag atcagcaact acaccaacac cgtgtacagg 1920  
ctgctggagg agagccagaa ccagcaggag aggaacgaga aggacctgt ggcctggac 1980  
agctggaaga acctgtggag ctggttcgac atcaccact ggctgtggta catcaagatc 2040  
ttcatcatca tcgtggccg cctgatcgcc ctgaggatca tttcgccgt gctgagcata 2100

gtgaacaggg tgaggcaggg ctacagcccc ctgagcttcc agaccctgac ccccaacccc 2160  
ggcgcccccg acaggctggg caggatcgag gaggagggcg gcaagcagga cagggacagg 2220  
agcatcaggc tggtaacgg cttcctggcc ctggcctggg acgacctgag gaacctgtgc 2280  
ctgttcagct accacaggct gagggacttc accctggtgg ccgcccagggt ggtggagctg 2340  
5 ctgggcagga acagcctgag gggcctgcag aggggctggg aggccctgaa gtacctggc 2400  
agcctggtgc agtactgggg ccaggagctg aagaagagca ccatcagcct ggtggacacc 2460  
atcgccatcg ccgtggccga gggcacccgac aggatcatcg agctggtgca gggcctgtgc 2520  
aggccatct acagcatccc caggaggatc aggcaaggct tcgaggccgc cctgcagtga 2580  
taaagatctc tcgaggagct caagc 2605